THE PROBLEM WITH SCIENCE

The Reproducibility Crisis and What To Do About It

R. BARKER BAUSELL

OXFORD

The Problem with Science

The Problem with Science

The Reproducibility Crisis and What to Do About It

R. BARKER BAUSELL

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

Library of Congress Cataloging-in-Publication Data
Names: Bausell, R. Barker, 1942– author.
Title: The problem with science : the reproducibility crisis and what to do about it / R. Barker Bausell, Ph.D.
Description: New York, NY : Oxford University Press, [2021] | Includes bibliographical references and index.
Identifiers: LCCN 2020030312 (print) | LCCN 2020030313 (ebook) | ISBN 9780197536537 (hardback) | ISBN 9780197536551 (epub) | ISBN 9780197536568
Subjects: LCSH: Science—Research—Methodology.
Classification: LCC Q126.9 .B38 2021 (print) | LCC Q126.9 (ebook) | DDC 507.2/1—dc23
LC record available at https://lccn.loc.gov/2020030312
LC ebook record available at https://lccn.loc.gov/2020030313

DOI: 10.1093/oso/9780197536537.001.0001

1 3 5 7 9 8 6 4 2

Printed by Sheridan Books, Inc., United States of America

Contents

III. STRATEGIES FOR INCREASING THE REPRODUCIBILITY OF PUBLISHED SCIENTIFIC RESULTS

A Brief Note

This book was written and peer reviewed by Oxford University Press before the news concerning the "problem" in Wuhan broke, hence no mention of COVID-19 appears in the text. Relatedly, since one of my earlier books, *Snake Oil Science: The Truth About Complementary and Alternative Medicine*, had been published by Oxford more than a decade ago, I had seen no need to pursue this line of inquiry further since the bulk of the evidence indicated that alternative medical therapies were little more than cleverly disguised placebos, with their positive scientific results having been facilitated by substandard experimental design, insufficient scientific training, questionable research practices, or worse. So, for this book, I chose to concentrate almost exclusively on a set of problems bedeviling mainstream science and the initiative based thereupon, one that has come to be called "the reproducibility crisis."

However, as everyone is painfully aware, in 2020, all hell broke loose. The internet lit up advocating bogus therapies; the leaders of the two most powerful countries in the world, Xi Jinping and Donald Trump, advocated traditional Chinese herbals and a household cleaner, respectively; and both disparaged or ignored actual scientific results that did not support their agendas. Both world leaders also personally employed (hence served as role models for many of their citizens) unproved, preventive remedies for COVID-19: traditional Chinese herbal compounds by Xi Jinping; hydroxychloroquine (which is accompanied by dangerous side effects) by Donald Trump.

This may actually be more understandable in Xi's case, since, of the two countries, China is undoubtedly the more problematic from the perspective of conducting and publishing its science. As only one example, 20 years ago, Andrew Vickers's systematic review team found that 100% of that country's alternative medical trials (in this case acupuncture) and 99% of its conventional medical counterparts published in China were positive. And unfortunately there is credible evidence that the abysmal methodological quality of Chinese herbal medical research itself (and not coincidentally the almost universally positive results touting their efficacy) has continued to this day.

To be fair, however, science as an institution is far from blameless in democracies such as the United States. Few scientists, including research methodologists such as myself, view attempting to educate our elected officials on scientific issues as part of their civic responsibility.

So while this book was written prior to the COVID-19 pandemic, there is little in it that is not relevant to research addressing future health crises such as this (e.g., the little-appreciated and somewhat counterintuitive [but well documented] fact that early findings in a new area of inquiry often tend to be either incorrect or to report significantly greater effect sizes than follow-up studies). It is therefore my hope that one of the ultimate effects of the reproducibility crisis (which again constitutes the subject matter of this book) will be to increase the societal utility of science as well as the public's trust therein. An aspiration that will not be realized without a substantial reduction in the prevalence of the many questionable research behaviors that permit and facilitate the inane tendency for scientists to *manufacture* (and publish) false-positive results.

Acknowledgments

First, I would like to thank the many conscientious and gifted researchers, methodologists, and statisticians whose insightful work informed this book. I have primarily depended upon their written word and sincerely hope that I have effectively captured the essence of their positions and research. I have not listed individual scientists in this acknowledgment since I have cited or been influenced by so many that I am reluctant to single out individuals for fear of omitting anyone (or including anyone who would prefer not to be so listed).

Finally, I would like to acknowledge my excellent Oxford University Press team who were extremely helpful and competent, and without whom the book would have never seen the light of day. Joan Bossert, Vice President/Editorial Director, for her support, who saw the promise in my original manuscript, selected very helpful peer reviewers, and guided me through the revision process. Phil Velinov, Assistant Editor, who seamlessly and competently coordinated the entire process. Suma George, Editorial Manager, who oversaw production. I would also like to extend my appreciation for my former editors at Oxford: Abby Gross and a special shout-out to the retired Marion Osmun—Editor Extraordinaire.

Introduction

This is a story about science. Not one describing great discoveries or the geniuses who make them, but one that describes the labors of scientists who are in the process of reforming the scientific enterprise itself. The impetus for this initiative involves a long-festering problem that potentially affects the usefulness and credibility of science itself.

The problem, which has come to be known as the *reproducibility crisis*, affects almost all of science, not one or two individual disciplines. Like its name, the problem revolves around the emerging realization that much—perhaps most—of the science being produced cannot be reproduced. And scientific findings that do not replicate are highly suspect if not worthless.

So, three of the most easily accomplished purposes of this book are

1. To present credible evidence, based on the published scientific record, that there exists (and has existed for some time) a serious reproducibility crisis that threatens many, if not most, sciences;
2. To present a menu of strategies and behaviors that, if adopted, have the potential of downgrading the problem from a crisis to a simple irritant; and
3. To serve as a resource to facilitate the teaching and acculturation of students aspiring to become scientists.

The book's potential audience includes

1. Practicing scientists who have not had the time or the opportunity to understand the extent of this crisis or how they can personally avoid producing (and potentially embarrassing) irreproducible results;
2. Aspiring scientists, such as graduate students and postdocs, for the same reasons;
3. Academic and funding administrators who play (whether they realize it or not) a key role in perpetuating the crisis; and

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0001

4. Members of the general public interested in scientific issues who are barraged almost daily with media reports of outrageously counterintuitive findings or ones that contradict previously ones.

Some readers may find descriptors such as "crisis" for an institution as sacrosanct as science a bit hyperbolic, but in truth this story has two themes. One involves a plethora of wrongness and one involves a chronicling of the labors of a growing cadre of scientists who have recognized the seriousness of the problem and have accordingly introduced evidence-based strategies for its amelioration.

However, regardless of semantic preferences, this book will present overwhelming evidence that a scientific crisis does indeed exist. In so doing it will not constitute a breathless exposé of disingenuous scientific blunders or bad behavior resulting in worthless research at the public expense. Certainly some such episodes compose an important part of the story, but, in its totality, this book is intended to educate as many readers as possible to a serious but addressable societal problem.

So, in a sense, this is an optimistic story representing the belief (and hope) that the culture of science itself is in the process of being altered to usher in an era in which (a) the social and behavioral sciences (hereafter referred to simply as the *social sciences*) will make more substantive, reproducible contributions to society; and (b) the health sciences will become even more productive than they have been in past decades. Of course the natural and physical sciences have their own set of problems, but only a handful of reproducibility issues from these disciplines have found their way into the present story since their methodologies tend to be quite different from the experimental and correlational approaches employed in the social and health sciences.

For the record, although hardly given to giddy optimism in many things scientific, I consider this astonishing 21st-century reproducibility awakening (or, in some cases, reawakening) to be deservedly labeled as a *paradigmatic shift* in the Kuhnian sense (1962). Not from the perspective of an earth-shattering change in scientific theories or worldviews such as ushered in by Copernicus, Newton, or Einstein, but rather in a dramatic shift (or change) in the manner in which scientific research is *conducted* and *reported*. These are behavioral and procedural changes that may also redirect scientific priorities and goals from a cultural emphasis on *publishing* as many professional articles as humanly possible to one of ensuring that what is published is *correct*, *reproducible*, and hence has a chance of being at least potentially useful.

However, change (whether paradigmatic or simply behavioral) cannot be fully understood or appreciated without at least a brief mention of what it replaces. So permit me the conceit of a very brief review of an important methodological initiative that occurred in the previous century.

The Age of Internal and External Validity

For the social sciences, our story is perhaps best begun in 1962, when a research methodologist (Donald T. Campbell) and a statistician (Julian C. Stanley) wrote a chapter in a handbook dealing with research on teaching of all things. The chapter garnered considerable attention at the time, and it soon became apparent that its precepts extended far beyond educational research. Accordingly, it was issued as an 84-page paperback monograph entitled *Experimental and Quasi-Experimental Designs for Research* (1966) and was promptly adopted as a supplemental textbook throughout the social sciences.

But while this little book's influence arguably marked the methodological coming of age for the social sciences, it was preceded (and undoubtedly was greatly influenced) by previous methodology textbooks such as Sir Ronald Fisher's *The Design of Experiments* (1935), written for agriculture researchers but influencing myriad other disciplines as well, and Sir Austin Bradford Hill's *Principles of Medical Statistics* (1937), which had an equally profound effect upon medical research.

The hallmark of Campbell and Stanley's remarkable little book involved the naming and explication of two constructs, *internal* and *external validity*, accompanied by a list of the research designs (or architecture) that addressed (or failed to address) the perceived shortcomings of research conducted in that era. Internal validity was defined in terms of whether or not an experimental outcome (generally presumed to be positive) was indeed a function of the intervention rather than extraneous events or procedural confounds. External validity addressed the question of:

> To what populations, settings, treatment variables, and measurement variables can this effect [presumably positive or negative] be generalized? (p. 5)

Of the two constructs, internal validity was the more explicitly described (and certainly the more easily addressed) by a list of 12 "threats"

thereto—most of which could be largely avoided by the random assignment of participants to experimental conditions. External validity, relevant only if internal validity was ensured, was so diffuse and expansive that it was basically given only lip service for much of the remainder of the century. Ironically, however, the primary arbiter of external validity (replication) also served as the same bottom line arbiter for the reproducible–irreproducible dichotomy that constitutes the basic subject matter of this book.

Campbell and Stanley's basic precepts, along with Jacob Cohen's (1977, 1988) seminal (but far too often ignored work) on statistical power, were subsequently included and cited in hundreds of subsequent research methods textbooks in just about every social science discipline. And, not coincidentally, these precepts influenced much of the veritable flood of methodological work occurring during the next several decades, not only in the social sciences but in the health sciences as well.

Unfortunately, this emphasis on the avoidance of structural (i.e., experimental design) at the expense of procedural (i.e., behavioral) confounds proved to be insufficient given the tacit assumption that if the architectural design of an experiment was reasonably sound and the data were properly analyzed, then any positive results accruing therefrom could be considered correct 95% of the time (i.e., the complement of the statistical significance criterion of $p \leq 0.05$). And while a vast literature did eventually accumulate around the avoidance of these procedural confounds, less attention was paid to the possibility that a veritable host of investigator-initiated questionable research practices might, purposefully or naïvely, artifactually produce false-positive, hence irreproducible, results.

From a scientific cultural perspective, this mindset was perhaps best characterized by the writings of Robert Merton (1973), a sociologist of science whose description of this culture would be taken as Pollyannaish satire if written today. In his most famous essay ("Science and the Social Order") he laid out "four sets of institutional imperatives—universalism, communism [sharing of information not the political designation], disinterestedness, and organized skepticism—[that] are taken to comprise the ethos of modern science" (p. 270).

Scientific ethos was further described as

> [t]he ethos of science is that affectively toned complex of values and norms which is held to be binding on the man of science. The norms are expressed in the form of prescriptions, proscriptions, preferences, and permissions.

> They are legitimatized in terms of institutional values. These imperatives, transmitted by precept and example and reinforced by sanctions, are in varying degrees internalized by the scientist, thus fashioning his scientific conscience or, if one prefers the latter-day phrase, his superego. (1973, p. 269, although the essay itself was first published in 1938)

While I am not fluent in Sociologese, I interpret this particular passage as describing the once popular notion that scientists' primary motivation was to discover truth rather than to produce a publishable p-value ≤ 0.05. Or that most scientists were so firmly enculturated into the "ethos" of their calling that any irreproducible results that might accrue were of little concern given the scientific process's "self-correcting" nature.

To be fair, Merton's essay was actually written in the 1930s and might have been somewhat more characteristic of science then than in the latter part of the 20th and early 21st centuries. But his vision of the cultural aspect of science was prevalent (and actually taught) during the same general period as were internal and external validity. Comforting thoughts certainly, but misconceptions that may explain why early warnings regarding irreproducibility were ignored.

Also in fairness, Merton's view of science was not patently incorrect: it was simply not sufficient. And the same can be said for Campbell and Stanley's focus on internal validity and the sound research designs that they fostered. Theirs might even qualify as an actual methodological paradigm for some disciplines, and it was certainly not incorrect. It was in fact quite useful. It simply was not sufficient to address an as yet unrecognized (or at least unappreciated) problem with the avalanche of scientific results that were in the process of being produced.

So while we owe a professional debt of gratitude to the previous generation of methodologists and their emphasis on the necessity of randomization and the use of appropriate designs capable of negating most experimental confounds, it is now past time to move on. For this approach has proved impotent in assuring the reproducibility of research findings. And although most researchers were aware that philosophers of science from Francis Bacon to Karl Popper had argued that a quintessential prerequisite for a scientific finding to be valid resides in its reproducibility (i.e., the ability of other scientists to replicate it), this crucial tenet was largely ignored in the social sciences (but taken much more seriously by physical scientists—possibly because they weren't required to recruit research

participants). Or perhaps it was simply due to their several millennia experiential head start.

In any event, ignoring the reproducibility of a scientific finding is a crucial failing because research that is not reproducible is worthless and, even worse, is detrimental to its parent science by (a) impeding the accumulation of knowledge, (b) squandering increasingly scarce societal resources, and (c) wasting the most precious of other scientists' resources—*their* time and ability to make *their* contributions to science. All failings, incidentally, that the reproducibility initiative is designed to ameliorate.

Another Purpose of This Book

While this book is designed to tell a scientific story, to provide practicing scientists with a menu of strategies to adopt (and behaviors to avoid) for assuring that *their* research can be reproduced by other scientists, or even to serve as a resource for the teaching of reproducibility concepts and strategies to aspiring scientists, these are not the ultimate purposes of the reproducibility initiative—hence not mine either. For while knowledge and altruistic motives may have some traction with those contemplating a career in science or those desiring to change their current practices, such resolutions face powerful competition in the forms of career advancement, families, and the seductive charms of direct deposit.

The ultimate purpose of the myriad dedicated methodologists whose work is described herein involves a far more ambitious task: the introduction of a *cultural* change in science itself to one that demands not only the avoidance of behaviors specifically designed to produce positive results, but also the adoption of a number of strategies that require additional effort and time on the part of already stressed and overworked scientists—a culture dedicated to the production of *correct* inferences to the extent that John Ioannidis (2005) will someday be able to write a rebuttal to his pejorative (but probably accurate) subhead from "Most Research Findings Are False for Most Research Designs and for Most Fields" to "False Positive Results Have Largely Disappeared from the Scientific Literatures." A culture in which the most potent personal motivations are not to produce hundreds of research publications or garner millions of dollars in research funding but to contribute knowledge to their scientific discipline that was previously unknown to anyone, anywhere. And conversely, a culture in which (short of actual fraud) the most embarrassing

professional incident that can occur for a scientist is for his or her research to be declared irreproducible due to avoidable questionable research practices when other scientists attempt to replicate it.

The Book's Plan for a Very Small Contribution to This Most Immodest Objective

Almost everyone is aware of what Robert Burns and John Steinbach had to say about the plans of mice and men, but planning is still necessary even if its objective is unattainable or nonexistent. So the book will begin with the past and present conditions that facilitate the troubling prevalence of irreproducible findings in the scientific literature (primarily the odd fact that many disciplines almost exclusively publish positive results in preference to negative ones). Next a *very* brief (and decidedly nontechnical) overview of the role that p-values and statistical power play in reproducibility/irreproducibility along with one of the most iconic modeling exercises in the history of science. The next several chapters delineate the behavioral *causes* (i.e., questionable research practices [QRPs]) of irreproducibility (accompanied by suggested solutions thereto) followed by a few examples of actual scientific pathology which also contribute to the problem (although hopefully not substantially). Only then will the replication process itself (the ultimate arbiter of reproducibility) be discussed in detail along with a growing number of very impressive initiatives dedicated to its widespread implementation. This will be followed by equally almost impressive initiatives for improving the publishing process (which include the enforcement of preregistration and data-sharing requirements that directly impact the reproducibility of what is published). The final chapter is basically a brief addendum positing alternate futures for the reproducibility movement, along with a few thoughts on the role of education in facilitating the production of reproducible results and the avoidance of irreproducible ones.

The Sciences Involved

While the book is designed to be multidisciplinary in nature, as mentioned previously it unavoidably concentrates on the social and health sciences. An inevitable emphasis is placed on psychological research since

that discipline's methodologists have unquestionably been leaders in (but by no means the only contributors to) reproducibility thought and the implementation of strategies designed to ameliorate the unsettling preponderance of false-positive results. However, examples from other sciences (and contributions from their practitioners) are provided, including laboratory and preclinical research (on which some of the most crucial human experimentation is often based) with even a nod or two to the physical sciences.

But while some of the book's content may seem irrelevant to practitioners and students of the purely biological and physical sciences, the majority of the key concepts discussed are relevant to almost all empirically based disciplines. Most sciences possess their own problems associated with publishing, the overproduction of positive results, statistical analysis, unrecognized (or hidden) questionable research practices, instrumental insensitivity, inadequate mentoring, and the sad possibility that there are just too many practicing scientists who are inadequately trained to ensure that their work is indeed reproducible.

A Few Unavoidable Irritants

Naturally, some of the content will be presented in more detail than some will prefer or require, but all adult readers have had ample practice in skipping over content they're either conversant with or uninterested in. To facilitate that process, most chapters are relatively self-contained, with cursory warnings of their content posted at the conclusion of their immediately preceding chapter.

Also, while the story being presented almost exclusively involves the published work of others, I cannot in good consciousness avoid inserting my own opinions regarding this work and the issues involved. I have, however, attempted to clearly separate my opinions from those of others.

Otherwise, every topic discussed is supported by credible empirical evidence, and every recommendation tendered is similarly supported by either evidence or reasoned opinions by well-recognized reproducibility thinkers. This strategy has unavoidably necessitated a plethora of citations which only constitute a mere fraction of the literature reviewed. For readability purposes, this winnowing process has admittedly resulted in an unsystematic review of cited sources, although the intent was to represent an overall consensus of

those thinkers and researchers who have contributed to this crucial scientific movement.

A Very Little About My Perspective

In my long academic and statistical consulting career, I have personally witnessed examples of pretty much all of the "good," "bad," and "ugly" of scientific practices. Following a brief stint as an educational researcher, my writing and research has largely focused on the methodology of conducting research and the statistical analysis of its results. I even once published an annotated guide to 2,600 published methodological sources encompassing 78 topics, 224 journals, and 125 publishers (Bausell, 1991). The tome was dedicated "to the three generations of research methodologists whose work this book partially represents" (p. viii).

In the three decades that followed, the methodological literature virtually exploded with the publication of more articles, topic areas, and journals (and, of course, blogs) than in the entire history of science prior to that time. And while this work has been extremely beneficial to the scientific enterprise, its main contribution may have been the facilitation of the emergence of a new generation of methodologists studying (and advocating for) the reproducibility of scientific results.

Naturally, as a chronicler, I could hardly avoid recognizing the revolutionary importance of this latter work, not just to research methodology but also to the entire scientific enterprise. My primary motivation for telling this story is to hopefully help promulgate and explain the importance of its message to the potential audiences previously described. And, of course, I dedicate the book "to the present generation of reproducibility methodologists it partially represents."

I must also acknowledge my debt to three virtual mentors who have guided me in interpreting and evaluating scientific evidence over the past two decades, philosophers of science from the recent and distant past whose best-known precepts I have struggled (not always successfully) to apply to the subject matter of this book as well. In chronological order these individuals are

1. William of Occam, the sternest and most fearsome of my mentors, whose most important precept was the *parsimony principle,* which

can be reduced to embracing the least involved explanation for the occurrence of a phenomenon that both fits the supporting data and requires the fewest assumptions (and usually constitutes the simplest explanation);

2. Yogi of Bronx, one of whose more important precepts was the principle of prediction, succinctly stated as "It's tough to make predictions, especially about the future"; and
3. Robert Park, the title of whose most important book pretty much speaks for itself: *Voodoo Science: The Road from Foolishness to Fraud (2000)*. And, as a disclaimer, Bob (the only one of my three mentors I ever actually met) never suggested that mainstream science had arrived at this latter destination—only that far too many scientists have blissfully traveled down that road.

And Finally an Affective Note

Some of the articles, quotations therefrom, and even some of my comments thereupon may appear overly critical. (Although for my part I have made a serious attempt to avoid doing so via the book's many revisions and greatly facilitated by one of its anonymous peer reviewer's very helpful comments.) It is important to remember, however, that what we are dealing with here is a paradigmatic shift (or radical change for those who prefer something less pompous) in the way in which science is conducted and published so *none of us should be overly judgmental*. It takes time for entrenched behaviors to change, and most practicing researchers (including myself) have violated one or more of the movement's precepts in the past, often unthinkingly because we hadn't been aware (or taught) that some of the questionable research behaviors discussed here were actually contraindicated.

After all, it wasn't until sometime around 2011–2012 that the scientific community's consciousness was bombarded with irreproducibility warnings via the work of scientists such as those discussed in this book (although warning shots had been fired earlier by scientists such as Anthony Greenwald in 1975 and John Ioannidis in 2005). However, this doesn't mean that we, as formal or informal peer reviewers, should be all that forgiving going forward regarding obviously contradicted practices such as failures to preregister protocols or flagrantly inflating p-values. We should just be civil in doing so.

So Where to Begin?

Let's begin with the odd phenomenon called *publication bias* with which everyone is familiar although many may not realize either the extent of its astonishing prevalence or its virulence as a facilitator of irreproducibility.

References

Bausell, R. B. (1991). *Advanced research methodology: An annotated guide to sources*. Metuchen, NJ: Scarecrow Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cohen, J. (1977, 1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Fisher, R. A. (1935). *The design of experiments*. London: Oliver & Boyd.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Hill, A. B. (1935). *Principles of medical statistics*. London: Lancet.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. (N. W. Storer, Ed.). Chicago: University of Chicago Press.

Park, R. (2000). *Voodoo science: the road from foolishness to fraud*. New York: Oxford University Press.

PART I

BACKGROUND AND FACILITATORS OF THE CRISIS

1
Publication Bias

Whether we refer to the subject matter of our story as a crisis or a paradigmatic shift, it must begin with a consideration of a primary facilitator of irreproducibility, one that has come to be called *publication bias*. This phenomenon can be defined most succinctly and nonpejoratively as *a tendency for positive results to be overrepresented in the published literature*—or as "the phenomenon of an experiment's results [here amended to "any study's results" not just experiments] determining its likelihood of publication, often over-representing positive findings" (Korevaar, Hooft, & ter Riet, 2011).

Surprisingly, the artifact's history actually precedes the rise of the journal system as we know it. Kay Dickersin (1991), a methodologist who specialized among other things in biases affecting the conduct of systematic reviews, provides a very brief and fascinating history of publication bias, suggesting that it was first referred to at least as early as the 17th century by Robert Boyle, often referred to as "one of the founders of modern chemistry, and one of the pioneers of modern experimental scientific method" (https://en.wikipedia.org/wiki/Robert_Boyle).

According to Dr. Dickerson, Boyle "was credited [in 1680] with being the first to report the details of his experiments and the precautions necessary for their *replication* [italics added because this process will soon become an integral part of our story]" (p. 1385). And even earlier, long before the rise of scientific journals, "Boyle lamented in 1661 that scientists did not write up single results but felt compelled to refrain from publishing until they had a 'system' worked out that they deemed worthy of formal presentation" (p. 1386). Or, in Boyle's words,

> But the worst inconvenience of all is yet to be mentioned, and that is, that whilst this vanity of thinking men obliged to write either systems or nothing is in request, many excellent notions or experiments are, by sober and modest men, suppressed." (p. 1386)

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0002

But, as we all know, over time some things improve while others get even worse. And the latter appears to be the case with publication bias, which has been germinating at least since the mid-20th century and remains in full bloom at the end of the second decades of the 21st century.

By Way of Illustration, a 20th-Century Parable

Suppose a completely fictitious first-year educational research graduate student's introduction to research came in his first methods class taught by a similarly hypothetical senior educational researcher (actually trained as a psychologist) more than a half-century ago. The lecture might have gone something like this:

> Whatever research you do on whatever topic you choose to do it on, always ensure that you can find at least one statistically significant p-value to report. *Otherwise you won't get published,* and you're going to need at least five publications per year if you want to have a halfway decent career. But since your hypothesis may be wrong, always hedge your bet by (a) employing a control group that you know *should* be inferior to your intervention, (b) including several different variables which can be substituted as your outcome if necessary, and (c) gleaning as much information from the student records (e.g., standardized tests, past grades—recalling that he was referring educational research) as possible which can be employed as covariates or blocking variables.

Now it is not known (or remembered) whether this long ago, hypothetical graduate student considered this to be most excellent advice or to reside somewhere on the continuum between absurd and demented. If the latter, he would have soon learned the hard way that the absence of a statistically significant p-value or two did indeed greatly reduce the probability of obtaining a cherished "publish with minor revisions" letter from a journal editor. (The "p" of course denotes probability and an obtained p-value ≤ 0.05 had, for many scientists and their sciences, become synonymous with statistical significance, the correctness of a tested hypothesis, and/or the existence of a true effect.)

In any event much of the remainder of the lecture and the course was given over to avoiding the aforementioned (Campbell & Stanley, 1966) threats to

internal validity, with a nod to its external counterpart and an occasional foray into statistical analysis. What was not typically considered (or taught) in those days was the possibility that false-positive results were clogging scientific literatures by unreported (and often unconsidered) strategies perfectly designed to produce publishable positive results—*even when participants were randomly assigned to conditions*. After all, empirical results could always be replicated by other researchers if they were sufficiently interested—which of course very few then were (or even are today).

So, thus trained and acculturated, our long-forgotten (or hypothetical) graduate student conducted many, many educational experiments and obtained statistical significance with some. And those he failed to do so he didn't bother to submit for publication once he had discovered for himself that his instructor had been right all along (i.e., that their rejection rate was several times greater than those blessed with p-values ≤ 0.05 as opposed to those cursed with the dreaded $p > 0.05$). After all, why even bother to write up such "failures" when the time could be more profitably spent conducting more "successful" studies?

Given his ambition and easy access to undergraduate and high school research participants he might have even conducted a series of 22 experiments in a failed effort to produce a study skill capable of producing more learning from a prose passage than simply reading and rereading said passage for the same amount of time—a hypothetical debacle that might have resulting in 22 statistically nonsignificant differences (none of which was ever submitted for publication). And once he might have even had the audacity to mention the quest itself at a brown bag departmental luncheon, which resulted in vigorous criticisms from all sides for wasting participants' time conducting nonsignificant studies.

So the Point of This Obviously Fictitious Parable Is?

Well certainly no one today would conduct such an absurd program of research for no other reason than the creation of an effective learning strategy. So let's just employ it to consider the ramifications of *not* publishing statistically nonsignificant research: some of which include the following.

First, it has come to be realized that it is unethical to ask people to volunteer to participate in research and not make the results available to other scientists. Or, as one group of investigators (Krzyzanowska, Pintilie, &

Tannock, 2003) more succinctly (and bluntly) state, "Nonpublication breaks the contract that investigators make with trial participants, funding agencies, and ethics boards" (p. 496).

Second, regardless of whether or not human participants are employed, publishing negative results

1. Permits other scientists from avoiding dead end paths that may be unproductive.
2. Potentially provides other scientists with an idea for a more effective intervention or a more relevant outcome variable (in the hypothetical scenario, this might have involved using *engagement* in the study activity as the outcome variable, thereby allowing the participants to take as much time as they needed to master the content), and
3. Encourages the creation of useful rather than spurious theories (the latter of which are more likely to be corroborated by positive results in the complete absence of negative ones).

But disregarding hypothetical scenarios or non-hypothetical ethical concerns, the reluctance to publish research associated with p-values > 0.05 has an even more insidious consequence. Decades upon decades of *publication bias* have resulted in a plethora of false-positive results characterizing the literatures of many entire scientific disciplines. And this, in turn, has become a major contributor to the lack of reproducibility in these disciplines since anyone with a sufficiently accomplished "skill set" can support almost any theory or practice.

The Relationship Between Publication and Scientific Irreproducibility

While irreproducibility and publication bias are inextricably linked, their relationship isn't necessarily causal since neither is a necessary nor sufficient condition of the other. For example, there are legitimate reasons why any *given* negative study may not be published, such as its authors' legitimate belief that the study in question was methodologically flawed, underpowered, and/or simply too poorly conducted to merit publication. (Solely for convenience, a "negative study or finding" will be referred to henceforth as one that explicitly or implicitly hypothesizes the occurrence of a statistically

significant result, but a nonsignificant one is obtained.) Or, far less commonly, an investigator may hypothesis the *equivalence* between experimental interventions, but these studies require specialized analytic approaches (see Bausell, 2015) and are relatively rare outside of clinical (most commonly pharmaceutical) trials.

Some unacceptable (but perhaps understandable) reasons for investigators' not attempting to publish a negative finding could include

1. Their misremembrance of their introductory statistics professors' cautionary edicts against overinterpreting a negative finding (or their misinterpretation of the precept that it is impossible to "prove" a null hypothesis),
2. Embarrassment resulting from the failure to support a cherished hypothesis that they remain convinced was correct, or, as with our hypothetical student,
3. Their understanding of the difficulties of publishing a negative study leading to a decision to spend their time conducting "positive" research rather than "wasting" time writing up "failed" studies.

When queried via surveys of investigators (e.g., Greenwald, 1975; Cooper, DeNeve, & Charlton, 1997; Weber, Callaham, & Wears, 1998), other reasons listed include (a) lack of time, (b) the belief that a negative study won't be published (as will be discussed shortly, there is definitely an editorial bias against doing so), (c) loss of interest, (d) a realization that the study was too flawed or underpowered (more on this in Chapter 2), and/or (e) the fact that the study was never intended to be published in the first place (presumably because it was a pilot study or was designed to test a single element of a larger study [e.g., the reliability or sensitivity of an outcome variable]). More recently a survey of laboratory animal researchers (ter Riet, Korevaar, Leenaars, et al., 2012) arrived at similar findings regarding reasons for not publishing negative studies and concluded that investigators, their supervisors, peer reviewers, and journal editors all bear a portion of the blame for the practice.

In a sense it is not as important to delineate the contributors to the scarcity of negative studies in the published literature as it is to identify the causes and ameliorate the consequences of the high prevalence of their false-positive counterparts. But before proceeding to that discussion, let's briefly consider a sampling of the evidence supporting the existence and extent of publication bias as well as the number of disciplines affected.

The Prevalence of Positive, Published Results in Science as a Whole

Undoubtedly the most ambitious effort to estimate the extent to which positive results dominate the scientific literature was employed by Daniele Fanelli, who contrasted entire sciences on the acceptance or rejection of their stated hypothesis. In his first paper (2010), 2,434 studies published from 2000 to 2007 were selected from 10,837 journals in order to compare 20 different scientific fields with respect to their rate of positive findings.

The clear "winner" turned out to be psychology-psychiatry, with a 91.5% statistical significance rate although perhaps the most shocking findings emanating from this study were that (a) all 20 of the sciences (which basically constitute the backbone of our species' empirical, inferential scientific effort) reported positive published success rates of greater than 70%, (b) the average positive rate for the 2,434 studies was 84%, and (c) when the 20 sciences were collapsed into three commonly employed categories all obtained positive rates in excess of 80% (i.e., biological sciences = 81%, physical sciences = 84%, and social sciences = 88%).

In his follow-up analysis, Fanelli (2011) added studies from 1990 to 1999 to the 2000 to 2007 sample just discussed in order to determine if these positive rates were constant or if they changed over time. He found that, as a collective, the 20 sciences had witnessed a 22% increase in positive findings over this relatively brief time period. Eight disciplines (clinical medicine, economic and business, geoscience, immunology, molecular biology-genetics, neuroscience-behavior, psychology-psychiatry, and pharmacology-toxicology) actually reported positive results at least 90% of the time by 2007, followed by seven (agriculture, microbiology, materials science, neuroscience-behavior, plants-animals, physics, and the social sciences) enjoying positive rates of from 80% to 90%. (Note that since the author did not report the exact percentages for these disciplines, these values were estimated based on figure 2 of the 2011 report.)

Now, of course, neither of these analyses is completely free of potential flaws (as are none of the other 35 or so studies cited later in this chapter), but they constitute the best evidence we have regarding the prevalence of publication bias. (For example, both studies employed the presence of a key sentence, "test*the hypothes*," in abstracts only, and some disciplines did not rely on p-values for their hypothesis tests.) However, another investigator (Pautasso, 2010) provides a degree of confirmatory evidence for the

Fanelli results by finding similar (but somewhat less dramatic) increases in the overall proportion of positive results *over time* using (a) four different databases, (b) different key search phrases ("no significant difference/s" or "no statistically significant difference/s"), (c) different disciplinary breakdowns, and, for some years, (d) only titles rather than abstracts.

A Brief Disclaimer Regarding Data Mining

Textual data mining meta-research (aka meta-science) is not without some very real epistemological limitations. In the studies cited in this book, a finite number of words or phrases (often only one) is employed which may not capture all of the studies relevant to the investigators' purposes or may be irrelevant to some articles. These choices may inadvertently introduce either overestimates or underestimates of the prevalence of the targeted behavior or phenomena. And, of course, it is a rare meta-research study (which can be most succinctly defined as research on research and which includes a large proportion of the studies discussed in this book) that employs a causal design employing randomization of participants or other entities to groups.

With that said, these studies constitute much of the best evidence we have for the phenomena of interest. And data mining efforts are becoming ever more sophisticated, as witnessed by the Menke, Roelandse, Ozyurt, and colleagues (2020) study discussed in Chapter 11 in which multiple search terms keyed to multiple research guidelines were employed to track changes in reproducibility practices over time.

Other Studies and Approaches to the Documentation of Publication Bias

One of the more commonly employed methods of studying the prevalence of positive findings in published literatures involves examining actual p-values in specific journals rather than the data mining approaches just discussed. One of the earliest of these efforts was conducted by Professor T. D. Sterling (1959) via a survey of four general psychology journals which found that an astonishing 97% of the studies published therein rejected the null hypothesis (i.e., achieved statistical significance at the 0.05 level or below). A little over a decade later, the team of Bozarth and Roberts (1972) found similar results

(94%), as did Sterling and two colleagues (Sterling, Rosenbaum, & Weinkam, 1995) two decades or so later than that. (This time around, Sterling and his co-investigators searched eight psychology journals and found the prevalence of positive results within one or two percentage points [96%] of the previous two efforts.) Their rather anticlimactic conclusion: "These results also indicate that practices leading to publication bias have not changed over a period of 30 years" (p. 108). However, as the previous section illustrated, it has changed for the worst more recently in some disciplines.

While surveys of journal articles published in specific journals constitute the earliest approach to studying publication bias, more recently, examinations of meta-analyses (both with respect to the individual studies comprising them and the meta-analyses themselves) have become more popular vehicles to explore the phenomenon. However, perhaps the most methodologically sound approach involves comparisons of the publication status of positive and negative longitudinal trials based on (a) institutional review board (IRB) and institutional animal care and use committee (IACUC) applications and (b) conference abstracts. Two excellent interdisciplinary reviews of such studies (Song, Parekh-Bhurke, Hooper, et al., 2009; Dwan, Altman, Arnaiz, et al., 2013) found, perhaps unsurprisingly by now, that positive studies were significantly more likely to be published than their negative counterparts.

More entertaining, there are even *experimental* documentations of the phenomenon in which methodologically oriented investigators, with the blessings of the journals involved, send out two almost identical versions of the same bogus article to journal reviewers. "Almost identical" because one version reports a statistically significant result while the other reports no statistical significance. The positive version tended to be significantly more likely to be (a) accepted for publication (Atkinson, Furlong, & Wampold, 1982) and (b) rated more highly on various factors such as methodological soundness (Mahoney, 1977), or (c) both (Emerson, Warme, Wolf, et al., 2010).

A Quick Recap

While psychology may be the scientific leader in reporting positive results, Fanelli (2011) and Pautasso (2010) have demonstrated that the

phenomenon is by no means found only in that discipline. In fact the majority of human and animal empirical studies employing p-values as the means for accepting or rejecting their hypotheses seem to be afflicted with this particular bias. In support of this rather pejorative generalization, the remainder of this chapter is given over to the presence of publication bias in a sampling of (a) *subdisciplines* or research *topics within disciplines* and (b) the methodological factors known to be subject to (or associated with) publication bias.

But First, the First of Many Caveats

First, the literature on publication bias is too vast and diverse to be reviewed either exhaustively or systematically here. Second, there is no good reason to do so since Diogenes, even with a state-of-the-art meta-science lantern, would have difficulty finding any methodologically oriented scientist who is not already aware of the existence of publication bias or who does not consider it be a significant scientific problem. And finally, the diversity of strategies designed to document publication bias makes comparisons or overall summaries of point estimates uninterpretable. Some of these strategies (which have been or will be mentioned) include (a) specific journal searches, (b) meta-research studies involving large databases, (c) meta-science examinations of reported p-values or key words associated with statistical significance/nonsignificance, (d) meta-analyses employing funnel plots to identify overabundancies of small studies reporting large effects, and (e) longitudinal follow-ups of studies from conference abstracts and IRB proposals to their ensuing journal publication/nonpublication.

An Incomplete Sampling of the Topic Areas Affected

- Psychology research (e.g., historically: Sterling, 1959; Atkinson et al., 1982; Sterling et al., 1995, and too many others to mention)
- Cancer clinical trials (Berlin, Begg, & Louis, 1989; Krzyzanowska et al., 2003)
- Cancer studies in general (De Bellefeuille, Morrison, & Tannock, 1992)
- Animal studies (Tsilidis, Panagiotou, Sena, et al., 2013)

- Child health (Hartling, Craig, & Russell, 2004)
- Medical randomized clinical trials (RCTs; Dickersin, Chan, & Chalmers, 1987)
- Preclinical stroke (Sena, van der Worp, Bath, et al., 2010)
- Psychotherapy for depression (Cuijpers, Smit, Bohlmeijer, et al., 2010; Flint, Cuijpers, & Horder, 2015)
- Pediatric research (Klassen, Wiebe, Russell, et al., 2002)
- Gastroenterology research (Timmer et al., 2002) and gastroenterological research cancer risk (Shaheen, Crosby, Bozymski, & Sandler, 2000)
- Antidepressant medications (Turner, Matthews, Linardatos, et al., 2008)
- Alternative medicine (Vickers, Goyal, Harland, & Rees, 1998; Pittler, Abbot, Harkness, & Ernst, 2000)
- Obesity research (Allison, Faith, & Gorman, 1996)
- Functional magnetic resonance imaging (fMRI) studies of emotion, personality, and social cognition (Vul, Harris, Winkielman, & Pashler, 2009) plus functional fMRI studies in general (Carp, 2012)
- Empirical sociology (Gerber & Malhotra, 2008)
- Anesthesiology (De Oliveira, Chang, Kendall, et al. 2012)
- Political behavior (Gerber, Malhotra, Dowling, & Doherty, 2010)
- Neuroimaging (Ioannidis, 2011; Jennings & Van Horn, 2012).
- Cancer prognostic markers (Kyzas, Denaxa-Kyza, & Ioannidis, 2007; Macleod, Michie, Roberts, et al., 2014)
- Education (Lipsey & Wilson, 1993; Hattie, 2009)
- Empirical economics (Doucouliagos, 2005)
- Brain volume abnormalities (Ioannidis, 2011)
- Reproductive medicine (Polyzos, Valachis, Patavoukas, et al., 2011)
- Cognitive sciences (Ioannidis, Munafò, Fusar-Poli, et al., 2014)
- Orthodontics (Koletsi, Karagianni, Pandis, et al., 2009)
- Chinese genetic epidemiology (Pan, Trikalinos, Kavvoura, et al., 2005)
- Drug addiction (Vecchi, Belleudi, Amato, et al., 2009)
- Biology (Csada, James, & Espie, 1996)
- Genetic epidemiology (Agema, Jukema, Zwinderman, & van der Wall, 2002)
- Phase III cancer trials published in high-impact journals (Tang, Pond, Welsh, & Chen, 2014)

Plus a Sampling of Additional Factors Associated with Publication Bias

- Multiple publications more so than single publication of the same data (Tramèr, Reynolds, Moore, & McQuay, 1997; Schein & Paladugu, 2001); although Melander, Ahlqvist-Rastad, Meijer, and Beermann (2003) found the opposite relationship for a set of Swedish studies
- The first hypothesis tested in multiple hypothesis studies less than single-hypothesis studies (Fanelli, 2010)
- Higher impact journals more so than low-impact journals (Tang et al., 2014, for cancer studies); but exceptions exist, such as the *Journal of the American Medical Association* (JAMA) and the *New England Journal of Medicine* (NEJM), for clinical trials (Olson, Rennie, Cook, et al., 2002)
- Non-English more so than English-language publications (Vickers et al., 1998; Jüni, Holenstein, Sterne, et al., 2003)
- RCTs with larger sample sizes less so than RCTs with smaller ones (Easterbrook, Berlin, Gopalan, & Matthews, 1991)
- RCTs less often than observational studies (e.g., epidemiological research), laboratory-based experimental studies, and nonrandomized trials (Easterbrook et al., 1991; Tricco, Tetzaff, Pham, et al., 2009)
- Research reported in complementary and alternative medicine journals more so than most other types of journals (Ernst & Pittler, 1997)
- Methodologically sound alternative medicine trials in high impact jounrals less so than their methodologically unsound counterparts in the same journals (Bausell, 2009)
- Meta-analyses more so than subsequent large RCTs on same topic (LeLorier, Gregoire, Benhaddad, et al, 1997).
- Earlier studies more so than later studies on same topic (Jennings & Van Horn, 2012; Ioannidis, 2008)
- Preregistration less often than no registration of trials (Kaplan & Irvin, 2015)
- Physical sciences (81%) less often than biological sciences (84%), less than social sciences (88%) (Fanelli, 2010), although Fanelli and Ioannidis (2013) found that the United States may be a greater culprit in the increased rate of positive findings in the "soft" (e.g., social) sciences than other countries
- Investigators reporting no financial conflict of interest less often than those who do have such a conflict (Bekelman, Li, & Gross, 2003;

Friedman & Richter, 2004; Perlis, Perlis, Wu, et al., 2005; Okike, Kocher, Mehlman, & Bhandari, 2007); all high-impact medical (as do most other medical) journals require a statement by all authors regarding conflict of interest.

- Pulmonary and allergy trials funded by pharmaceutical companies more so than similar trials funded by other sources (Liss, 2006)
- Fewer reports of harm in stroke research in published than non-published studies, as well as publication bias in general (Liebeskind, Kidwell, Sayre, & Saver, 2006)
- Superior results for prevention and criminology intervention trials when evaluated by the program developers versus independent evaluators (Eisner, 2009)
- And, of course, studies conducted by investigators known to have committed fraud or misconduct more so than those not so identified; it is a rare armed robbery that involves donating rather than stealing money.

A Dissenting Voice

Not included in this list is a systematic review (Dubben & Beck-Bornholdt, 2005) whose title, "Systematic Review of Publication Bias in Studies on Publication Bias," might appear to be parodic if it weren't for the fact that there is now a systematic review or meta-analysis available for practically every scientific topic imaginable. And, as should come as no surprise, the vast majority of meta-analytic conclusions are positive since the vast majority of their published literatures suffer from publication bias. (One meta-analytic database, the Cochrane Database of Systematic Reviews, is largely spared from this seemingly slanderous statement when the review involves the efficacy of a specific hypothesis involving a specific outcome; see Tricco et al., 2009.)

The Bottom Line

As a gestalt, the evidence is compelling that publication bias exists, buttressed by supporting (a) studies conducted over a period of decades, (b) the diversity

by which the evidence was produced, and (c) the number of disciplines involved. In addition, the Fanelli (2011) and Pautasso (2010) analyses indicate that the phenomenon has been operating for a considerable amount of time and appears to be actually accelerating in most disciplines (although this is close to numerically impossible for some areas whose percentage of positive results approach 100%).

Whether the large and growing preponderance of these positive results in the published scientific literatures is a cause, a symptom, or simply a facilitator of irreproducibility doesn't particularly matter. What is important is that the lopsided availability of positive results (at the expense of negative ones) distorts our understanding of the world we live in as well as retards the accumulation of the type of knowledge that science is designed to provide.

This phenomenon, considering (a) the sheer number of scientific publications now being produced (estimated to be in excess of 2 million per year; National Science Board, 2018) and (b) the inconvenient fact that most of these publications are positive, leads to the following rather disturbing implications:

1. Even if the majority of these positive effects are not the product of questionable research practices specifically designed to produce positive findings, and
2. If an unknown number of *correct* negative effects are not published, then
3. Other investigative teams (unaware of these unpublished studies) will test these hypotheses (which in reality are false) until someone produces a positive result by chance alone—or some other artifact, which
4. Will naturally be published (given that it is positive) even if far more definitive contradictory evidence exists—therefore contributing to an already error-prone scientific literature.

Unfortunately, some very persuasive evidence will be presented in the next few chapters to suggest that the *majority of this welter of positive scientific findings being published today (and published in the past) does not represent even marginally sound scientific findings*, but instead are much more likely to be categorically false. And if this is actually true (which, again, the evidence soon to be presented suggests), then we most definitely have a scientific crisis of epic proportions on our hands.

But What's to Be Done About Publication Bias?

First, given the ubiquitous nature (and long history) of the problem, perhaps some of the strategies that probably won't be particularly helpful to reduce publication bias should be listed. Leaning heavily (but not entirely) on a paper entitled "Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability" (Nosek, Spies, & Motyl, 2012), some of these ineffective (or at least mostly ineffective) strategies are

1. *Educational campaigns emphasizing the importance of publishing nonsignificant results as well as statistically significant ones*. While principled education is always a laudable enterprise, writing yet another article extolling the virtues of changing investigator, editorial, or student behaviors is unlikely be especially effective. There have been enough of such articles published over the past half-century or so.
2. *Creating journals devoted to publishing nonsignificant results*. Historically this strategy hasn't been particularly successful in attracting enough submission to become viable. Examples, old and new, include *Negative Results in Biomedicine*, *Journal of Negative Observations in Genetic Oncology*, *Journal of Pharmaceutical Negative Results* (this one is definitely doomed), *Journal of Articles in Support of the Null Hypothesis*, *The All Results Journals*, *New Negatives in Plant Science*, *PLoS ONE's Positively Negative Collection*, *Preclinical Reproducibility and Robustness Gateway*, and probably many others that have come and gone as some of these already have.
3. *Devoting dedicated space to publishing negative results in traditional journals*. Several journals have adopted this strategy and while it isn't a bad idea, it is yet to make any significant impact on the problem.
4. *Putting the burden on peer reviewers to detect false-positive results*. As editor-in-chief of an evaluation journal for 33 years, I can attest to the impossibility of this one. Providing reviewers with a checklist of potential risk factors (e.g., low power, the presence of uncharacteristically large effect sizes, counterintuitive covariates, p-values between 0.045 and 0.0499) might be helpful.
5. *Appealing to the ethical concerns of a new generation of scientists*. Such as, for example, campaigning to designate using human or animal participants in one's research and failing to publish the research

as unethical and a form of actual scientific misconduct (Chalmers, 1990).

Somehow we need to implement a cultural change in the sciences by convincing new and beginning investigators that publishing negative studies is simply a cost of doing business. Certainly it is understandable that conserving time and effort may be a personal priority but *everyone* associated with the publication process bears a responsibility for failing to correct the bias against publishing methodologically sound negative studies. And this list includes

1. Investigators, such as our hypothetical graduate student who slipped his 22 negative studies into Robert Rosenthal's allegorical "file drawer" (1979), never to be translated into an actual research report or accompanied by a sufficiently comprehensive workflow to do so in the future. (See Chapter 9 for more details of the concept of keeping detailed workflows, along with Phillip Bourne's [2010] discussion thereof.) For even though many of us (perhaps even our mythical graduate student) may have good intentions to publish all of our negative studies in the future, in time the intricate details of conducting even a simple experiment fade in the face of constant competition for space in long-term memory. And although we think we will, we never seem to have more time available in the future than we do now in the present.
2. Journal editors who feel pressure (personal and/or corporate) to ensure a competitive citation rate for their journals and firmly believe that publishing negative studies will interfere with this goal as well as reduce their readership. (To my knowledge there is little or no empirical foundation for this belief.) We might attempt to convince editors of major journals to impose a specified percentage annual limit on the publication of positive results, perhaps beginning as high as 85% and gradually decreasing it over time.
3. Peer reviewers with a bias against studies with p-values > 0.05. As mentioned previously, this artifact has been documented experimentally several times by randomly assigning journal reviewers to review one of two identical versions of a fake manuscript, with the exception that one reports statistical significance while the other reports nonsignificance. Perhaps mandatory peer review seminars and/or checklists could be

developed for graduate students and postdocs to reduce this peer reviewer bias in the future.

4. Research funders who would much rather report that they spent their money demonstrating that something works or exists versus that it does not. And since the vast majority of investigators' funding proposals hypothesize (hence promise) positive results, they tend to be in no hurry to rush their negative results into publication—at least until their next grant proposal is approved.
5. The public and the press that it serves are also human, with the same proclivities, although with an added bias toward the "man bites dog" phenomenon.

However, there are steps that all researchers can take to reduce the untoward effects of publication bias by such as

1. Immediately writing and *quickly* submitting negative studies for publication. And, if rejected, quickly resubmitting them to another journal until they are accepted (and they eventually will be given sufficient persistent coupled with the absolute glut of journals in most fields). And, as for any study, (a) transparently reporting any glitches in their conduct (which peer reviewers will appreciate and many will actually reward) and, (b) perhaps especially importantly for negative studies, explaining why the study results are important contributions to science;
2. Presenting negative results at a conference (which does not preclude subsequent journal publication). In the presence of insufficient funds to attend one, perhaps a co-author or colleague could be persuaded to do so at one that he or she plans to attend;
3. Serving as a peer reviewer (a time-consuming, underappreciated duty all scientists must perform) or journal editor (even worse): evaluating studies based on their design, conceptualization, and conduct rather than their results;
4. And, perhaps most promising of all, utilizing preprint archives such as the arXiv, bioRxiv, engrXiv, MedRxiv, MetaArXiv, PeerJ, PsyArXiv, SocArXiv, and SSRN, which do not discriminate against negative results. This process could become a major mechanism for increasing the visibility and availability of nonsignificant results since no peer review is required and there is no page limit, so manuscripts can be as long or as brief as their authors' desire.

An Effective Vaccine

One reason publication bias is so difficult to prevent involves its disparate list of "culprits" (e.g., investigators, journal editors, peer reviewers) and the unassailable fact that negative studies are simply very difficult to get published. One very creative preventive measure was designated as the *Registered Reports process*; it was championed by Chris Chambers and presumably first adopted by the journal *Cortex* and announced via a guest editorial by Brian Nosek and Daniël Lakens in 2014.

This innovative approach to publishing actually *prevents* reviewers and journal editors from discriminating against nonsignificant results and hence greatly incentivizes investigators to submit them for publication in the first place. Perhaps most comprehensively and succinctly described by Nosek, Ebersole, DeHaven, and Mellor (2018), Registered Reports and their accompanying preregistration are potentially one of the most effective strategies yet developed for preventing false-positive results due to publication bias. The process, in the authors' words, occurs as follows:

> With Registered Reports, authors submit their research question and methodology to the journal for peer review before observing the outcomes of the research. If reviewers agree that the question is sufficiently important and the methodology to test it is of sufficiently high quality, then the paper is given in-principle acceptance. The researchers then carry out the study and submit the final report to the journal. At second-stage review, reviewers do not evaluate the perceived importance of the outcomes. Rather, they evaluate the quality of study execution and adherence to the preregistered plan. In addition to the benefits of preregistration, this workflow addresses selective reporting of results and facilitates improving research designs during the peer review process. (p. 2605)

Apparently the Registered Reports initiative is taking hold since, as early as 2018, Hardwicke and Ioannidis found that 91 journals had adopted the procedure across a number of scientific disciplines. A total of 109 Registered Reports had been published in psychology, the majority of which were replication studies (in which case they are referred to as "Registered Replication Reports" which will be discussed in more detail in Chapter 7).

In addition to its potential salutary effects on publication bias, the authors list several other important advantages of the Registered Report process resulting from the a priori peer review process. These include

1. The potential for methodological flaws being identified in time to avoid poorly designed studies being added to the literature;
2. Publishing judgments being based on the merits of the research question and design, rather than on the aesthetic characteristics of the findings;
3. Research being more likely to be comprehensively and transparently reported since authors know they can be held accountable based on the original registered protocol; and
4. Questionable research practices (which will be discussed in detail in Chapter 3) will be greatly reduced since the original protocol can (and hopefully will) be compared with the final research report by journal editors and reviewers.

And as yet another reason for publishing negative results (in the long run they are more likely to be correct than their positive counterparts): Poynard, Munteanu, Ratziu, and colleagues (2002) conducted a completely unique analysis (at least as far I can ascertain) with the possible exception of a book tangentially related thereto (Arbesman, 2012) entitled *The Half-Life of Facts: Why Everything We Know Has an Expiration Date*). The authors accomplished this by collecting cirrhosis and hepatitis articles and meta-analyses conducted between 1945 and 1999 in order to determine which of the original conclusions were still considered "true" by 2000. Their primary results were

> Of 474 [resulting] conclusions (60%) were still considered to be true, 91 (19%) were considered to be obsolete, and 98 (21%) were considered to be false. The half-life of truth was 45 years. The 20-year survival of conclusions derived from meta-analysis was lower (57% ± 10%) than that from nonrandomized studies (87% ± 2%) ($p < 0.001$) or randomized trials (85% ± 3%) ($p < 0.001$). (p. 888)

However, one of the study's subgroup analyses (which should always be interpreted with caution) was of even more interest to us here since it could be interpreted as a comparison between positive and negative findings

with respect to their production of false-positive versus false-negative conclusions. Namely that, "in randomized trials, the 50-year survival rate was higher for 52 negative conclusions (68% ± 13%) than for 118 positive conclusions (14% ± 4%, $p < 0.001$)" (p. 891).

Thus, while based on a single study (and a secondary analysis at that), this finding could constitute a case for the scientific importance of not preferring positive studies to negative studies since the latter may be more likely to be true in the long run.

So What's Next?

Before beginning the exploration of irreproducibility's heart of darkness (and reproducibility's bright potential), it may be helpful to quickly review three introductory statistical constructs (statistical significance, statistical power, and the effect size). These are key adjuncts to reproducibility and touch just about everything empirical in one way or another. However, since everyone who has suffered through an introductory statistics course (which probably encompasses 99% of the readers of this book), a natural question resides in why the repetition is necessary.

The answer, based on my not so limited experience, is that many researchers, graduate students, and postdocs do not have a deep understanding of these concepts and their ubiquitous applications to all things statistical. And even for those who do, the first few pages of Chapter 2 may be instructive since statistical significance and power play a much more important (perhaps *the* most important) role in reproducibility than is commonly realized. The p-value, because it is so easily *gamed* (as illustrated in Chapter 3), is, some would say, set too high to begin with, while statistical power is so often ignored (and consequently set too low).

References

Agema, W. R., Jukema, J. W., Zwinderman, A. H., & van der Wall, E. E. (2002). A meta-analysis of the angiotensin-converting enzyme gene polymorphism and restenosis after percutaneous transluminal coronary revascularization: Evidence for publication bias. *American Heart Journal, 144*, 760–768.

Allison, D. B., Faith, M. S., & Gorman, B. S. (1996). Publication bias in obesity treatment trials? *International Journal of Obesity and Related Metabolic Disorders, 20*, 931–937.

Arbesman, S. (2012). *The half-life of facts: Why everything we know has an expiration date*. New York: Penguin.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance reviewer evaluations, and the scientific process: Is there a statistically significant relationship? *Journal of Counseling Psychology, 29*, 189–194.

Bausell, R. B. (2009). Are positive alternative medical therapy trials credible? Evidence from four high-impact medical journals. *Evaluation & the Health Professions, 32*, 349–369.

Bausell, R. B. (2015). *The design and conduct of meaningful experiments involving human participants: 25 scientific principles*. New York: Oxford University Press.

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *Journal of the American Medical Association, 289*, 454–465.

Berlin, J. A., Begg, C. B., & Louis, T. A. (1989). An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association, 84*, 381–392.

Bourne, P. E. (2010). What do I want from the publisher of the future? *PLoS Computational Biology, 6*, e1000787.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist, 27*, 774–775.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage, 63*, 289–300.

Chalmers, I. (1990). Underreporting research is scientific misconduct. *Journal of the American Medical Association, 263*, 1405–1408.

Cooper, H. M., DeNeve, K. M., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2*, 447–452.

Csada, R. D., James, P. C., & Espie, R. H. M. (1996). The "file drawer problem" of non-significant results: Does it apply to biological research? *Oikos, 76*, 591–593.

Cuijpers, P., Smit, F., Bohlmeijer, E., et al. (2010). Efficacy of cognitive-behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *British Journal of Psychiatry, 196*, 173–178.

De Bellefeuille, C., Morrison, C. A., & Tannock, I. F. (1992). The fate of abstracts submitted to a cancer meeting: Factors which influence presentation and subsequent publication. *Annals of Oncology, 3*, 187–191.

De Oliveira, G. S., Jr., Chang, R., Kendall, M. C., et al. (2012). Publication bias in the anesthesiology literature. *Anesthesia & Analgesia, 114*, 1042–1048.

Dickersin, K. (1991). The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association, 263*, 1385–1389.

Dickersin, K., Chan, S., Chalmers, T. C., et al. (1987). Publication bias and clinical trials. *Controlled Clinical Trials, 8*, 343–353.

Doucouliagos, C. (2005). Publication bias in the economic freedom and economic growth literature. *Journal of Economic Surveys, 19*, 367–387.

Dubben, H-H., & Beck-Bornholdt, H-P. (2005). Systematic review of publication bias in studies on publication bias. *British Medical Journal, 331*, 433–434.

Dwan, K., Altman, D. G., Arnaiz, J. A., et al. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias: An updated review. *PLoS ONE, 8*, e66844.

Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet, 337*, 867–872.

Eisner, M. (2009). No effects in independent prevention trials: Can we reject the cynical view? *Journal of Experimental Criminology, 5*, 163–183.

Emerson. G. B., Warme, W. J., Wolf, F. M., et al. (2010). Testing for the presence of positive-outcome bias in peer review. *Archives of Internal Medicine, 170*, 1934–1939.

Ernst, E., & Pittler, M. H. (1997). Alternative therapy bias. *Nature, 385*, 480.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5*, e10068.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904.

Fanelli, D., & Ioannidis, J. P. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of the Sciences, 110*, 15031–15036.

Flint, J., Cuijpers, P., & Horder, J. (2015). Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychological Medicine, 45*, 439–446.

Friedman, L. S., & Richter, E. D. (2004). Relationship between conflicts of interest and research results. *Journal of General Internal Medicine, 19*, 51–56.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods and Research, 37*, 3–30.

Gerber, A. S., Malhotra, N., Dowling, C. M., & Doherty, D. (2010). Publication bias in two political behavior literatures. *American Politics Research, 38*, 591–613.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Hardwicke, T., & Ioannidis, J. (2018). Mapping the universe of registered reports. *Nature Human Behaviour, 2*, 10.1038/s41562–018–0444–y.

Hartling, L., Craig, W. R., & Russell, K. (2004). Factors influencing the publication of randomized controlled trials in child health research. *Archives of Adolescent Medicine, 158*, 984–987.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Ioannidis, J. P. (2011). Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry, 68*, 773–780.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology 19*, 640–648.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., et al. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Science, 19*, 235–241.

Jennings, R. G., & Van Horn, J. D. (2012). Publication bias in neuroimaging research: Implications for meta-analyses. *Neuroinformatics, 10*, 67–80.

Jüni, P., Holenstein, F., Sterne, J., et al. (2003). Direction and impact of language bias in meta-analysis of controlled trials: Empirical study. *International Journal of Epidemiology, 31*, 115–123.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS ONE, 10*, e0132382.

Klassen, T. P., Wiebe, N., Russell, K., et al. (2002). Abstracts of randomized controlled trials presented at the Society for Pediatric Research Meeting. *Archives of Pediatric and Adolescent Medicine, 156*, 474–479.

Koletsi, D., Karagianni, A., Pandis, N., et al. (2009). Are studies reporting significant results more likely to be published? *American Journal of Orthodontics and Dentofacial Orthopedics, 136*, 632e1–632e5.

Korevaar, D. A., Hooft, L., & ter Riet (2011). Systematic reviews and meta-analyses of preclinical studies: Publication bias in laboratory animal experiments. *Laboratory Animals, 45*, 225–230.

Krzyzanowska, M. K., Pintilie, M., & Tannock, I. F. (2003). Factors associated with failure to publish large randomized trials presented at an oncology meeting. *Journal of the American Medical Association, 290*, 495–501.

Kyzas, P. A., Denaxa-Kyza, D., & Ioannidis, J. P. (2007). Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer, 43*, 2559–2579.

LeLorier, J., Gregoire, G., Benhaddad, A., et al. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine, 337*, 536–542.

Liebeskind, D. S., Kidwell, C. S., Sayre, J. W., & Saver, J. L. (2006). Evidence of publication bias in reporting acute stroke clinical trials. *Neurology, 67*, 973–979.

Lipsey, M. W., & Wilson, D. B. (1993). Educational and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

Liss, H. (2006). Publication bias in the pulmonary/allergy literature: Effect of pharmaceutical company sponsorship. *Israeli Medical Association Journal, 8*, 451–544.

Macleod, M. R., Michie, S., Roberts, I., et al. (2014). Increasing value and reducing waste in biomedical research regulation and management. *Lancet*, *383*, 176–185.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*, 161–175.

Melander, H., Ahlqvist-Rastad, J., Meijer, G., & Beermann, B. (2003). Evidence b(i)ased medicine-selective reporting from studies sponsored by pharmaceutical industry: Review of studies in new drug applications. *British Medical Journal, 326*, 1171–1173.

Menke, J., Roelandse, M., Ozyurt, B., et al. (2020). Rigor and Transparency Index, a new metric of quality for assessing biological and medical science methods. bioRxiv http://doi.org/dkg6;2020

National Science Board. (2018). *Science and engineering indicators 2018.* NSB-2018-1. Alexandria, VA: National Science Foundation. https://www.nsf.gov/statistics/indicators/.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600–2606.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives in Psychological Science, 7*, 615–631.

Okike, K., Kocher, M. S., Mehlman, C. T., & Bhandari, M. (2007). Conflict of interest in orthopaedic research: An association between findings and funding in scientific presentations. *Journal of Bone and Joint Surgery, 89*, 608–613.

Olson, C. M., Rennie, D., Cook, D., et al. (2002). Publication bias in editorial decision making. *Journal of the American Medical Association, 287*, 2825–2828.

Pan, Z., Trikalinos, T. A., Kavvoura, F. K., et al. (2005). Local literature bias in genetic epidemiology: An empirical evaluation of the Chinese literature [see comment]. *PLoS Medicine, 2,* e334.

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics, 85*, 193–202.

Perlis, R. H., Perlis, C. S., Wu, Y., et al. (2005). Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *American Journal of Psychiatry, 162*, 1957–1960.

Pittler, M. H., Abbot, N. C., Harkness, E. F., & Ernst, E. (2000). Location bias in controlled clinical trials of complementary/alternative therapies. *Journal of Clinical Epidemiology, 53*, 485–489.

Polyzos, N. P., Valachis, A., Patavoukas, E., et al. (2011). Publication bias in reproductive medicine: From the European Society of Human Reproduction and Embryology annual meeting to publication. *Human Reproduction, 26*, 1371–1376.

Poynard, T., Munteanu, M., Ratziu, V., et al. (2002). Truth survival in clinical research: An evidence-based requiem? *Annuals of Internal Medicine, 136*, 888–895.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Schein, M., & Paladugu, R. (2001). Redundant surgical publications: Tip of the iceberg? *Surgery, 129*, 655–661.

Sena, E. S., van der Worp, H. B., Bath, P. M., et al. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biology, 8*, e1000344.

Shaheen, N. J., Crosby, M. A., Bozymski, E. M., & Sandler, R. S. (2000). Is there publication bias in the reporting of cancer risk in Barrett's esophagus? *Gastroenterology, 119*, 333–338.

Song, F., Parekh-Bhurke, S., Hooper, L., et al. (2009). Extent of publication bias in different categories of research cohorts: A meta-analysis of empirical studies. *BMC Medical Research Methodology, 9,* 79.

Sterling, T. D. (1959). Publication decision and the possible effects on inferences drawn from tests of significance-or vice versa. *Journal of the American Statistical Association, 54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49*, 108–112.

Tang, P. A., Pond, G. R., Welsh, S., & Chen, E. X. (2014). Factors associated with publication of randomized phase III cancer trials in journals with a high impact factor. *Current Oncology, 21*, e564–572.

ter Riet, G., Korevaar, D. A., Leenaars, M., et al. (2012). Publication bias in laboratory animal research: A survey on magnitude, drivers, consequences and potential solutions. *PLoS ONE, 1*(9), e43404,

Timmer, A., Hilsden, R. J., Cole, J., et al. (2002). Publication bias in gastroenterological research: A retrospective cohort study based on abstracts submitted to a scientific meeting. *BMC Medical Research Methodology*, *2*, 7.

Tramèr, M. R., Reynolds, D. J., Moore, R. A., & McQuay, H. J. (1997). Impact of covert duplicate publication on meta-analysis: A case study. *British Medical Journal, 315*, 635–640.

Tricco, A. C., Tetzaff, J., Pham, B., et al. (2009). Non-Cochrane vs. Cochrane reviews were twice as likely to have positive conclusion statements: Cross-sectional study. *Journal of Clinical Epidemiology, 62*, 380–386.

Tsilidis, K. K., Panagiotou, O. A., Sena, E. S., et al. (2013). Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS Biology, 11*, e1001609.

Turner, E. H., Matthews, A. M., Linardatos, E., et al. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252–260.

Vecchi, S., Belleudi, V., Amato, L., et al. (2009). Does direction of results of abstracts submitted to scientific conferences on drug addiction predict full publication? *BMC Medical Research Methodology, 9,* 23.

Vickers, A., Goyal, N., Harland, R., & Rees, R. (1998). Do certain countries produce only positive results? A systematic review of controlled trials. *Controlled Clinical Trials, 19*, 159–166.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

Weber, E. J., Callaham, M. L., & Wears, R. L. (1998). Unpublished research from a medical specialty meeting: Why investigators fail to publish. *Journal of the American Medical Association, 280*, 257–259.

2

False-Positive Results and a Nontechnical Overview of Their Modeling

Let's begin with a definition of the most worrisome manifestation (and primary constituent) of irreproducible research findings: false-positive results. A *false-positive result* occurs when (a) an impotent intervention results in a statistically significant change in a specific outcome or (b) a relationship between two (or more) *unrelated* variables is found to be statistically significant. Or, more succinctly, it is a positive statistically significant result that cannot be reliably replicated. Or, more pejoratively, it is a positive finding that has resulted from (a) one or more egregious procedural/statistical errors or (b) investigator ignorance, bias, or fraud.

Of course, false-negative results are also problematic, but, given the high prevalence of publication bias in most disciplines, they are considerably more rare and won't be emphasized here. This doesn't mean that false negatives are unimportant since those emanating from clinical trials designed to test the efficacy of actually effective drugs or treatments could lead to tragic events. It is just that scenarios such as this appear to be quite rare in clinical research and almost nonexistent in the social science literatures.

There are three statistical constructs that contribute to the production of false-positive results.

1. *Statistical significance*, defined by the comparison between the probability level generated by a computer following the statistical analysis performed on study results (referred to here as the *p-value*) and the maximum probability level hypothesized by the investigator or based on a disciplinary consensus or tradition (referred to as the *alpha level*). If the obtained p-value is less than or exactly equal to (≤) the hypothesized or disciplinary conventional alpha level (typically 0.05), then statistical significance is declared.
2. *Statistical power* is most succinctly (and rather cavalierly) defined as the probability that a given study will result in statistical significance

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0003

(the minimum value of which is most often recommended to be set at 0.80). Statistical power is a function of (a) the targeted alpha level; (b) the study design; (c) the number of participants, animals, or other observations employed; and (d) our third statistical construct, the effect size.

3. The *effect size* is possibly the simplest of the three constructs to conceptualize *but without question* it is the most difficult of the three constructs to predict prior to conducting a study. It is most often predicted based on (a) a small-scale pilot study, (b) a review of the results of similar studies (e.g., meta-analyses), or (c) a disciplinary convention, which, in the social sciences, is often set at 0.50 based on Jacob Cohen's decades-old (1988) recommendation. Its prediction is also the most tenuous of the triad regardless of how it is generated. If the effect size is overestimated, even when a hypothesized effect actually exists, its attendant study will be more difficult to replicate without adjustments such as an increased sample size or the use of questionable research practices (QRPs). If the effect size is underestimated, replication is more likely (even in the absence of QRPs), and, if the true effect size under investigation is sufficiently large, the attendant study will most likely be either trivial or constitute a major scientific finding. Since this latter scenario occurs with extreme rarity and the overestimation of effect sizes is far more common—whether predicted a priori or based on study results—most of what follows will be based on this scenario.

All three constructs are based on a statistical model called the *normal* or *bell-shaped curve*, which is often depicted as shown in Figure 2.1.

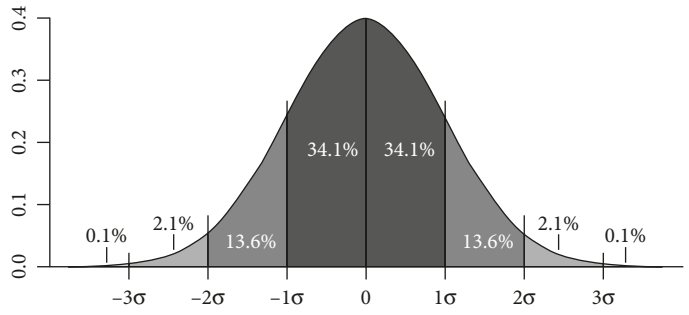

Figure 2.1 The bell-shaped, normal curve.
https://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg

All three are therefore integrally related to one another. For example, all else being equal:

1. The lower (or more stringently) the alpha level is set (i.e., the manner in which statistical significance is defined), the lower the power will be unless the study design is properly adjusted (e.g., increasing the number of participants or other observations [aka the sample size] to be employed) and/or the larger the hypothesized effect size must be;
2. The higher the desired power, the larger the required sample size and/or the larger the hypothesized effect size must be (if the alpha level is not adjusted); and, obviously,
3. The smaller the hypothesized effect size the less statistical power will be available (unless the sample size is increased and/or the alpha level is not adjusted).

Naturally, since all three of these statistical constructs are based on the normal curve they, too, are all subject to the same rather restrictive governing assumptions plus a few of their own. But the normal curve has proved to be a rather useful model which is surprising robust to minor violations of many of its assumptions.

So let's now begin the modeling of false-positive results based on the statistical models just discussed. As depicted in Table 2.1, this diagram has

Table 2.1 A modeling of the probabilities of the four possible study outcomes in the absence of systematic bias

	What is "actually" true	
	The hypothesis is correct	**The hypothesis is incorrect**
Possible study outcomes		
The hypothesis is confirmed *(obtained p-value = alpha level = .05)*	[a] Correct Finding *(p of occurrence = statistical power = .80)*	[b] False-Positive Result *(p of occurrence = alpha level = .05)*
The hypothesis is not confirmed *(obtained p-value > .05)*	[c] False-Negative Result] (*p of occurrence = 1-statistical power = .20*]	[d] Correct Finding *(p of occurrence = 1-alpha = .95)*

For present purposes, systematic bias involves questionable research practices (QRPs), whether consciously or unconsciously employed.

confounded students in untold numbers of introductory statistics and research methods books for decades, sometimes with no warning that it represents a most tenuous statistical model that has little or no practical applicability to actual scientific practice.

What could be simpler? Of course, we never know whether the original hypothesis is correct or not, but this is a model, after all, so we must make a few assumptions. And since we know that hypotheses are almost always confirmed in published studies, why even worry about unconfirmed hypotheses? And certainly a false-positive rate of 5% is nothing to worry about and definitely not indicative of a crisis of any sort. So let's just concentrate on the first row of cells.

But, alas, even here some key assumptions are missing governing the use of these concepts (along with the effect size, which does not appear but, as just discussed, is integrally related to statistical significance and power). Assumptions which, if violated, render the false-positive result projection practically useless except as a starting point for additional modeling.

For example, let's start with one of the assumptions related to either the obtained p-value (or alpha level from an a priori perspective).

> A single obtained p-value (or single event alpha level) normally applies to an a priori specified hypothesis involving a *single* result associated with a *single* outcome unless both are adjusted downward to compensate for multiple events or conditions. If no adjustment is made and statistical significance is obtained, the accruing result is more likely to be incorrect (hence a false-positive result) than in the presence of an appropriate adjustment.

To illustrate, let's assume that an investigator conducted a simple two-group experiment designed to ascertain if the two groups differed on a single outcome variable and obtained a p-value of 0.04999. However if *two* outcome variables were employed and the same p-value was obtained for both findings, neither result would actually be statistically significant following an appropriate downward adjustment of either the original alpha level or the obtained p-value. And, not coincidentally, the original power would have been an overestimate because its calculation involved an inflated p-value.

As another example, as mentioned earlier, power is sometimes inadequately defined as the probability of obtaining statistical significance but a more accurate (if unwieldy) definition must take the possibility of systematic error (or bias) into account such as, statistical power is

> the probability that an experiment [or any type of empirical study for that matter] will result in *statistical significance* if that significance level is appropriate for the design employed [i.e., is properly adjusted], if the study is properly conducted [i.e., in the absence of unavoidable glitches and QRPs], and if its hypothesized effect size is correct. (Bausell & Li, 2002, p. 14)

Many thoughtful reproducibility scholars would probably consider the simplistic model just discussed as too basic to even be considered a false-positive model at all. However, my reason for presenting it in this context is that it provides an opportunity to review some important points regarding the statistical significance and statistical power concepts that are often taken for granted. And, after all, it did provide a barebones model for predicting the prevalence of this most worrisome manifestation (and primary constituent) of irreproducible research findings.

So with these potential precursors to false-positive results at least partially discussed, let's now turn our attention to what is by far the most iconic, influential, and widely read article in the entire scientific reproducibility arena. Authored by John Ioannidis, it has garnered well over 8,000 citations and might just have provided the spark that ignited the reproducibility revolution itself and, not coincidentally, produced a far, far different probability estimate for cell "b" of Table 2.1 (which, it will be remembered, suggested a 5% occurrence of false-positive results in the absence of bias).

Why Most Published Research Findings Are False

John P. A. Ioannidis (2005)

This modeling effort is designed to estimate the rate of false-positive results. It employed the same statistical concepts as did Table 2.1 but added two totally new ones: *the ratio of true to no relationships in a discipline* and the even more important one regarding the likely effect of certain well-known biasing factors (e.g., conducting multiple analyses) known to inflate p-values and hence *produce* false-positive results when these inflated p-values are not corrected.

As an aside, statistical models such as this have important strengths and weaknesses which are bothersome to many scientists. However, they are pervasive in science as a whole, sometimes do not involve actual

observations of any sort, are not testable experimentally or observationally (at least when they are first made), and sometimes involve assumptions that may be completely demented but often constitute our only option to predict future events or estimate presently unobservable phenomena. Or, in George Box's much more eloquent and succinct aphorism: "All models are wrong, but some are useful."

But, returning to the model at hand, naturally the use of a *ratio of true to no relationships in a discipline or area of endeavor* immediately raises the question of how anyone could possibly estimate such a value since it is almost impossible to determine when even one research finding is categorically true (short of at least one rigorous replication). But that's the beauty (and utility) of theoretical models: they allow us to *assume* any (or as many different) values as we please for a presently unknowable construct.

The model itself employs a simple algebraic formula (perhaps first proposed by Wacholder et al., 2004) which can be used to ascertain the prevalence of false-positive results in an entire research arena. (Conversely, subtracting this value from 1.0 obviously provides the probability of the average scientific result being correct—hence a *true-positive* result.)

So while correctly estimating the number of true effects that actually exist in an entire scientific field is, to say the least, difficult, Professor Ioannidis wisely chose what appeared to be a realistic example from a field at the border of psychology and genetics—a field with a large literature designed to locate statistically significant correlations between various single-nucleotide polymorphisms (SNPs; of which there are an estimated 10 million) and various psychological constructs and diagnoses, such as general intelligence or schizophrenia (an actual, if unfortunate, empirical example of which will be presented shortly).

SNPs, pronounced "snips," constitute the primary source of genetic variations in humans and basically involve postconception changes in the sequence of a single DNA "building block." The overwhelming majority are benign mutations that have no known untoward or beneficial effects on the individual, but, in the proper location, they have the capacity to affect a gene's functions—one of which is increased susceptibility to a disease.

Ioannidis therefore chose this data mining arena as his example, assuming that 100,000 gene polymorphisms might be a reasonable estimate for the number of possible candidates for such an inquiry (i.e., the denominator of the required ratio), accompanied by a limited but defensible guess at the likely number of SNPs that might actually play a role (the

numerator) in the specific psychological attribute of interest. So, given these assumptions, let's suspend judgment and see where this exercise takes us.

For his imputed values, Ioannidis chose 0.05 for the significance criterion (customarily employed in the genomic field at that time but fortunately no longer), 0.60 for the amount of statistical power available for the analysis, and 10 for the number of polymorphisms likely to be associated with the attribute of interest, which Ioannidis hypothetically chose to be schizophrenia. (Dividing the best guess regarding the number of true relationships [10] by the number of analyses [100,000] yields the proportion of "true effects" in this hypothetical domain.)

Plugging these three values into the above-mentioned modeling formula produced an estimated false-positive rate above the 50% level and hence far above the 5% rate of false-positive results posited in Table 2.1. (And this, in turn, indicated that any obtained statistically significant relationship close to a p-value of 0.05 between a gene and the development of schizophrenia would probably be false.)

Ioannidis then went on to add two different scenarios to his model, both of which are known to increase the proportion of published false-positives results.

1. The rate of QRPs (i.e., systematic data analyses and/or investigator procedural practices that have been demonstrated to artifactually enhance the chances of producing statistical significant findings) and
2. A facet of publication bias in which 10 research teams are investigating the same topic but only 1 of the 10 finds statistically significant results (which of course means that this single positive finding would be considerably more likely to be published—*or even submitted for publication*—than would the other nine non-statistically significant results).

Not surprisingly the results of these additions to the model produced estimated false-positive percentages even higher than the original less restrictive model. So Ioannidis, not one to mince words, summarized his conclusions under the previously mentioned, very explicit subhead: "Most Research Findings Are False for Most Research Designs and for Most Fields"—which very nicely mirrors the article's equally pejorative title.

It is hard to say whether this subheading and the article's title are accurate or not, but certainly both are absolute "generalization whoppers." And, as is true for most models' assumptions, anyone can quibble with their real-world validity, including those employed in this study or even the almost universally accepted constructs of p-values, statistical power, and the normal curve.

However, since we *do* design and analyze our research based on these particular assumptions, it makes sense to use them in our efforts to estimate false-positive rates. So regardless of whether or not we agree with Professor Ioannidis's point estimates, we all owe him a debt for his modeling efforts and the following six corollaries he presents related to the genesis of false-positive results.

Hopefully, he will not object to these "corollaries" being repeating verbatim here or my attempts at succinct explanations for their mechanisms of action.

1. *The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.* Studies with small sample sizes are generally associated with less statistical power, but when such studies happen to generate positive results they are considerably more likely to be incorrect than their high-powered counterparts. However, since low-powered studies are more common than high-powered ones in most disciplines, this adds to these disciplines' false-positive rates. (As mentioned, low statistical power can also be a leading cause of false-negative results, but this is less problematic because so few negative studies are published.)
2. *The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.* As fields mature and the low hanging fruit has already been harvested, their effects become smaller and thus require increasingly larger samples to maintain acceptable statistical power. If these sample size compensations are not made accordingly, then power decreases and the rate of false positives increases. (And when effects move toward the limits of our instruments' capacity to reliably detect them, erroneous findings increase accordingly.)
3. *The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.* This corollary might have been stated more clearly, but what Ioannidis apparently means is that fields with lower pre-study

probabilities of being true (e.g., genetic association studies in which thousands upon thousands of relationships are tested and only a few true-positive effects exist) have a greater prevalence of false-positive results in comparison to fields in which fewer hypotheses are tested since said hypotheses must be informed by more and better preliminary, supportive data (e.g., large medical randomized controlled trials [RCTs], which are quite expensive to mount and must have preliminary data supporting their hypotheses before they are funded). In addition, clinical RCTs (perhaps with the exception of psychological and psychiatric trials) tend to be more methodologically sophisticated (e.g., via the use of double-blinded placebo designs) and regulated (e.g., the requirement that detailed protocols be preregistered, thereby decreasing the prevalence of a posteriori hypothesis changes). These conditions also reduce the prevalence of false-positive results.

4. *The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.* Here, the difference between publishing practices in high-impact, hypothesis-testing medical journals versus social science outlets is even greater than for the previous corollary. Efficacy studies such as those published in the *Journal of the American Medical Association* or the *New Journal of Medicine* customarily involve randomization of patients; a detailed diagram of patient recruitment, including dropouts; double blinding; veridical control groups (e.g., a placebo or an effective alternative treatment); a recognized health outcome (as opposed to idiosyncratic self-reported ones constructed by the investigators), pre-registration of the study protocol, including data analytic procedures; intent-to-treat analyses; and the other strategies listed in the Consolidated Standards of Reporting Trials (CONSORT) Statement of clinical medical trials (Schulz, Altman, & Moher for the Consort Group, 2010). Publication standards in the social sciences are far more "flexible" in these regards, most notably perhaps in the sheer number of self-reported idiosyncratic outcome variables that are quite distal from any recognized veridical social or behavioral outcome. More importantly the "greater flexibility" mentioned in this corollary also entails a greater prevalence of QRPs in the design and conduct of studies (see Chapter 3).

5. *The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.* Rather self-explanatory and encompasses self-interest, bias, fraud, and misconduct—all of which will be discussed in later chapters. An example of the biasing effects due to financial interests will be discussed with respect to pharmaceutical research in Chapter 10.
6. *The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.* If nothing else, "hotter" fields encourage publication bias such as via the author's genetic scenario in which the first team to achieve statistical significance is more likely to publish its results than the first team that finds no statistically significant effect.

And while Dr. Ioannidis' assumptions regarding genetic association studies (definitely a hot field) may appear unrealistic, an extremely impressive review of such studies involving a single genetic location and susceptibility to a specific disease suggests otherwise (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002). This latter team initially found 603 statistically significant findings associated with 268 individual genes that were associated with 603 statistically significant findings, 166 of which had been studied three or more times, and only six of these proved to be consistently replicable. So for those not keeping track, this makes Ioannidis's assertion that "most positive findings are false" appear rather modest.

A Second Modeling Exercise

Let's now consider a second modeling exercise that may have more applicability for social and behavioral experimentation with actual human participants. This model also requires an estimate regarding the prior probability of true effects and basically employs the same formula used by Ioannidis and proposed by Wacholder et al. However, since this one targets an entire discipline's experimentation rather than multiple analyses on the same dataset, its prior probability estimate may be somewhat of a greater stretch (but perhaps more applicable to experimental research).

Is the Replicability Crisis Overblown? Three Arguments Examined

Harold Pashler and Christine R. Harris (2012)

Targeting psychological experimentation as a whole, Pashler and Harris define the proportion of true-positive effects in their discipline as the percentage of effects that "researchers look for actually exist." They posit 10% as the most reasonable estimate (which is obviously a far cry from Ioannidis's 10^{-4} choice for genetic data monitoring). However, it may actually be an overestimate for disciplines such as educational research, which appears to lean heavily on trivial and repetitive hypotheses (Bausell, 2017) and therefore would presumably be excluded from Pashler and Harris's 10% estimate.

Inserting this 10% value into the formula along with the average power for psychological research (≈0.50) for detecting a titular alpha level of 0.05 (assuming an average effect size and a sample size of 20 participants per group in a two-group study) yielded a *discipline false-positive* estimate of 56% within the experimental psychology literature. Which, coincidentally (or not), is quite consonant with Ioannidis's "inflammatory" 2005 conclusion that "most research findings are false for most research designs and for most fields"—or at least for the field of experimental psychology.

Again, the primary weakness of this model resides in the choice of the imputed value for the prior probability of true effects (that "researchers look for actually exist"), which, of course, is subject to change over time. But since Harold Pashler and Christine Harris *are* well regarded psychological researchers, 10% is probably as reasonable an estimate as any.

However, as previously mentioned, the primary advantage of modeling resides in the ability to input as many different determinants of the endpoint of interest as the modeler desires. So, in this case, I have taken the liberty of expanding Pashler and Harris's illustrative results by adding a few additional values to the three input constructs in Table 2.2. Namely:

1. The prevalence of "true" disciplinary effects (.05 and .25 in addition to .10),
2. statistical power (.80 currently most often recommended and .50 to .35), and

Table 2.2 Estimation of false-positive results model

A: Proportion of discipline-wide studies assumed to have true effect	B: Power	C: Alpha (actual)	D: Proportion of false-positive results
.050	.800	.050	.54
.050	.500	.050	.66
.050	.350	.050	.73
.100	.800	.050	.36
.100	.500	.050	.47
.100	.350	.050	.56
.250	.800	.050	.16
.250	.500	.050	.23
.250	.350	.050	.30
.050	.800	.025	.37
.050	.500	.025	.49
.050	.350	.025	.58
.100	.800	.025	.22
.100	.500	.025	.31
.100	.350	.025	.39
.250	.800	.025	.09
.250	.500	.025	.13
.250	.350	.025	.18
.050	.800	.010	.19
.050	.500	.010	.28
.050	.350	.010	.35
.100	.800	.010	.10
.100	.500	.010	.15
.100	.350	.010	.20
.250	.800	.010	.04
.250	.500	.010	.06
.250	.350	.010	.08

Table 2.2 *Continued*

A: Proportion of discipline-wide studies assumed to have true effect	B: Power	C: Alpha (actual)	D: Proportion of false-positive results
.050	.800	.005	.11
.050	.500	.005	.16
.050	.350	.005	.21
.100	.800	.005	.05
.100	.500	.005	.08
.100	.350	.005	.11
.250	.800	.005	.02
.250	.500	.005	.03
.250	.350	.005	.04

3. in addition to the alpha level of .05 (.025, .01, and .005)—the latter, incidentally, being recommended by a consensus panel (Benjamin et al., 2017) for improving the reproducibility of new discoveries.

The operative component of this table is the final column (the proportion of positive results that are false), the values of which range from .73 to .02 (i.e., 73% to 2% of false positives in the psychological literature—or whatever that literature happens to be to which the inputted constructs might apply). That's obviously a huge discrepancy so let's examine the different assumptive inputs that have gone into this estimate (and it definitely is an estimate).

When the hypothesized proportion of true effects that scientists happen to be looking for ranges between 5% and 25% (the first column), the proportion of false-positive results (Column D) are powerfully affected by these hypothesized values. Thus if the discovery potential (Column A) is as low as .05, which might occur (among other possibilities) when scientists in the discipline are operating under completely fallacious paradigms, the average resulting proportion of false-positive results in the literature is .39 and ranges from .11 to .73. When true effects of .10 and .25 are assumed, the estimated published positive effects that are false drop to averages of .25 and .11, respectively.

Similarly as the average statistical power in a discipline increases from .35 to .80 the rate of published false-positive effects (irrespective of the alpha level and the estimated rate of false-positive effects in the literature) drops from .31 to .19. However it is the alpha level which is the most powerful independent determinant of false-positive results in this model. When the alpha is set at .05, the average rate of false-positive results in the literature averaged across the three levels of power and the four modeled assumed level of true effects is .46 or almost half of the published positive results in many scientific literatures. (And positive results, it will be recalled, comprise from .92 to .96 of psychology's published literature.)

An alpha level of .01 or .005 yields much more acceptable false-positive results (.16 and .09, respectively, averaged across the other two inputs). An alpha of .005, in fact, produces possibly acceptable false-positive results (i.e., < .20 or an average of .075) for all three projected rates of true effects and power levels of .80 and .50. Not coincidentally, a recent paper (Benjamin et al., 2017) in *Nature Human Behavior* (co-authored by a veritable Who's Who host of reproducibility experts) recommended that studies reporting *new discoveries* employ a p-value of .005 rather than .05. Ironically, a similar modeling conclusion was reached more than two decades ago in the classic article entitled "Effect Sizes and p Values: What Should Be Reported and What Should Be Replicated" (Greenwald, Gonzalez, Harris, & Guthrie, 1996).

For those interested in history, it could be argued that the first author, Anthony Greenwald, one of the 73 authors of the *Nature Human Behavior* paper just cited, foresaw the existence of the reproducibility crisis almost half a century ago in a classic paper "Consequences of Prejudice Against the Null Hypothesis" (1975). This paper also detailed what surely must have been the first (and, if not the first, surely the most creative) modeling demonstration of false-positive results so, just for the fun of it, let's briefly review that truly classic article.

Consequences of Prejudice Against the Null Hypothesis

Anthony G. Greenwald (1975)

To begin with, Greenwald sent a brief questionnaire to 48 authors and 47 reviewers of the *Journal of Personality and Social Psychology* querying them about practices regarding the alpha level, statistical power, and null results. (A quite acceptable 78% response rate was obtained.)

The results (recall that this survey was conducted more than four decades ago) of most interest to us today were as follows:

1. The mean probability level deemed most appropriate for rejecting a null hypothesis was .046 (quite close to the conventional .05 level).
2. The available statistical power deemed satisfactory for accepting the null hypothesis was .726 (also quite close to the standard .80 power recommendation for design purposes). Interestingly, however, only half of the sample responded to this latter query, and, based on other questions, the author concluded that only about 17% of the sample typically considered statistical power or the possibility of producing false-*negative* results prior to conducting their research. (In those days the primary effects of low power on reproducibility were not widely appreciated, and low power was considered problematic primarily in the *absence* of statistical significance.)
3. After conducting an initial full-scale test of the primary hypothesis and not achieving statistical significance, only 6% of the researchers said that they would submit the study without further data collection. (Hence publication bias is at least half a century old, as is possibly this QRP as well [i.e., presumably collecting additional data to achieve statistical significance without adjusting the alpha level].) A total of 56% said they would conduct a "modified" replication before deciding whether to submit, and 28% said they would give up on the problem. Only 10% said that they would conduct an exact replication.

Greenwald then used these and other questionnaire responses to model what he concluded to be a dysfunctional research publication system in which "there may be relatively few publications on problems for which the null hypothesis is (at least to a reasonable approximation) true, and of these, a high proportion will erroneously reject the null hypothesis" (p. 1). It is worth repeating that this prescient statement regarding reproducibility was issued almost a half-century ago and was, of course, largely ignored.

Greenwald therefore went on to conclude that by the time the entire process is completed the actual alpha level may have been raised from .05 to as high as .30 and the researcher "because of his investment in confirming his theory with a rejection of the null hypothesis, has overlooked the

possibility that the observed *x*-*y* relationship may be dependent on a specific manipulation, measure, experimenter, setting, or some combination of them" (p. 13). Then, under the heading "Some Epidemics of Type I Error" (aka false-positive results), he buttresses his case involving the inflation of alpha levels via several well-received past studies that found their way into textbooks but were later discarded because they couldn't be replicated.

So not only did Greenwald recognize the existence of a reproducibility crisis long before the announcement of the present one, he also (a) warned his profession about the problems associated with publication bias, (b) advocated the practice of replication at a time when it was even less common that it is today, and (c) provided guidelines for both the avoidance of publication bias and false-positive results. And while he is recognized to some extent for these accomplishments, he deserves an honored place in the pantheon of science itself (if there is such a thing).

So What Should We Make of These Modeling Efforts?

Are these models correct, incorrect, evidence of the existence of a true scientific crisis, or overblown empirical warnings that the sky is falling? No definitive answers exist for these questions because the questions themselves are inappropriate.

What these models provide, coupled with the incontrovertible existence of publication bias, is a clear warning that the published scientific literature may be characterized by far more false-positive results than either the scientific community or the public realize. But we've only started this story. We haven't even considered the calamitous effects of the primary villains of this story: a veritable host of QRPs and institutional impediments that foster the production of false-positive results. So let's consider some of these along with what appears to be an oxymoron: a downright amusing model illustrating the effects of QRPs on irreproducibility.

References

Bausell, R. B. (2017). *The science of the obvious: Education's repetitive search for what's already known*. Lanham, MD: Rowman & Littlefield.

Bausell, R. B., & Li, Y. F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical, and social sciences.* Cambridge: Cambridge University Press.

Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2017). Redefine statistical significance. *Nature Human Behavior, 2*, 6–10.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175–183.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine, 4*, 45–61.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–526.

Schulz, K. F., Altman, D. G., & Moher, D. for the Consort Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomized trials. *British Medical Journal, 340*, c332.

Wacholder, S., Chanock, S., Garcia-Closas M., et al. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, *96*, 434–442.

3

Questionable Research Practices (QRPs) and Their Devastating Scientific Effects

Publication bias is the most commonly ascribed cause and/or facilitator of the presumed high prevalence of false-positive results in the scientific literatures. However, in this case and facilitators are co-dependent with even more distal antecedents such as investigator ambitions, the need to provide for one's families, ignorance (caused by inadequate mentoring or day-dreaming in class), and undoubtedly several others that don't need to be identified.

So suffice it to say that while publication bias appears to be a facilitator of false-positive results, it is neither the only culprit nor probably the most important one. What are more impactful are an impressive set of investigative behaviors that conspire to produce both publication bias and false-positive results. And that dubious honor is reserved for questionable research practices (QRPs) that are almost entirely behavioral in nature.

In a sense the sheer number, diversity, and astonishingly high disciplinary prevalence of these miscreants constitute the primary linchpin of our story. What makes them even more problematic is the fact that the vast majority are never reported in the studies employing them. So our task in this chapter is to list and explicate these culprits. But first, a natural question presents itself.

Just How Prevalent Are These So-Called Contraindicated Practices?

The short answer is that we don't know, and there is really no completely accurate way of finding out. However there have been a number of surveys estimating the prevalence of both QRPs and outright fraudulent behaviors, and, like just about everything else scientific, these efforts have been the subject of at least one meta-analysis.

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0004

We owe this latter gift to Daniele Fanelli (2009), already introduced as an important and frequent contributor to the reproducibility literature, who has graced us with a meta-analysis involving 18 such surveys. Unfortunately the response rates of most of them were unimpressive (some would say inadequate), and, since they unavoidably involved self- or observed reports of antisocial behaviors, the results produced were undoubtedly underestimates of the prevalence of QRPs and/or fraud.

While the survey questions and sampling procedures in the studies reviewed by Dr. Fanelli were quite varied, the meta-analysis' general conclusions were as follows:

1. The percentage of self-reported actual *data fabrication* (the behavior universally deemed to constitute the worst example of fraudulent scientific practice) was low (approximately 2%),
2. Inevitably the respondents suspected (or had observed) higher percentages of misconduct among other scientists than they themselves had committed, and
3. "Once methodological differences were controlled for, cross study comparisons indicated that samples drawn exclusively from medical (including clinical and pharmacological) research reported misconduct more frequently than respondents in other fields or in mixed samples." (p. 10)

However, with respect to this third point, a more recent and unusually large survey (John, Loewenstein, & Prelec, 2012) casts some doubt thereupon and may have moved psychology to the head of the QRP class.

In this huge survey of almost 6,000 academic psychologists (of which 2,155 responded), questionnaires were emailed soliciting self-reported performance of 10 contraindicated practices known to bias research results. An intervention was also embedded within the survey designed to increase the validity of responses and permit modeling regarding the prevalence of the 10 targeted QRPs although, for present purposes, only the raw self-admission rates of these 10 QRPs listed in Table 3.1 will be discussed.

A cursory glance at these results indicates an unusually high prevalence of many of these practices, especially since they are based on self-reports. While a great deal of additional information was collected (e.g., respondents' opinions of the justifiability of these practices and whether their prevalence

Table 3.1 Self-admitted questionable research practices (QRPs) (for at least one episode) among academic psychologists

Questionable research practices	Prevalence
1. Failing to report all outcome variables	66.5%
2. Deciding whether to collect more data	58.0%
3. Selectively reporting studies that "worked"	50.0%
4. Deciding whether to exclude data following an interim analysis	43.4%
5. Reporting an unexpected finding as a predicted one	35.0%
6. Failing to report all study conditions	27.4%
7. Rounding off p-values (e.g., .054 to .05)	23.3%
8. Stopping study after desired results are obtained	22.5%
9. Falsely claiming that results are unaffected by demographic variables	4.5%
10. Falsifying data	1.7%

was greater in universities other than the respondents' [they were]), four of the author's conclusions stand out.

1. "Cases of clear scientific misconduct have received significant media attention recently, but less flagrantly *questionable research practices may be more prevalent and, ultimately, more damaging to the academic enterprise* [emphasis added]" (p. 524).
2. "Respondents considered these behaviors to be defensible when they engaged in them . . . but considered them indefensible overall" (p. 530) [scientific methodology evolves over time, which is one reason tolerance was previously recommended for some past sins—even the one attributed to our hypothetical graduate student].
3. "All three prevalence measures [which also included modeled QRP rates plus a follow-up survey of respondents] point to the same conclusion: a surprisingly high percentage of psychologists admit to having engaged in QRPs" (p. 530).
4. And most poignantly: "QRPs can waste researchers' time and stall scientific progress, as researchers fruitlessly pursue extensions of effects that are not real and hence cannot be replicated. More generally, the prevalence of QRPs raises questions about the credibility of research findings and threatens research integrity by producing unrealistically elegant results that may be difficult to match without engaging in such practices oneself. This can lead to a 'race to the bottom,' with questionable research begetting even more questionable research" (p. 531).

However, disciplinary differences in the prevalence of QRP practice (and especially when confounded by self-reports vs. the observance of others) are difficult to assess and not particularly important. For example, in the year between the Fanelli and the John et al. publications, Bedeian, Taylor, and Miller (2010) conducted a smaller survey of graduate business school faculty's *observance* of colleagues' committing 11 QRPs during the year previous to their taking the survey. This likewise produced a very high prevalence of several extremely serious QRPs, with the fabrication of data being higher in this survey (26.8%) than any I have yet encountered. Recall, however, that this and the following behaviors are reports of others' (not the respondents') behaviors:

1. A posteriori hypothesizing (92%),
2. Not reporting some methodological details or results (79%),
3. Selectively reporting data that supported the investigators' hypotheses (78%), and
4. Using ideas without permission or giving due credit (70%).

Now, of course, surveys are near the bottom of most triangles of evidence if they even make the list in the first place, and we shouldn't put too much stock in their point estimates given response rates and other biases inherent in self-reports. In addition, making comparisons between surveys is also challenging since (a) even questionnaires addressing the same topics usually contain slightly different items or item wordings, (b) the response instructions often differ (e.g., time period covered or self- vs. observed behaviors), and (c) the classification of behaviors differs (e.g., scientific fraud vs. QRPs vs. misconduct).

In way of illustration, while they would probably agree with some of the conclusions of the large John et al. survey, Drs. Fiedler and Schwarz (2015) argue (buttressed by a survey of their own) that the former overestimated the prevalence of QRPs due to ambiguities in wording and the fact that "prevalence" of a behavior cannot be calculated or inferred from the proportions of people who engaged in these behaviors only once. As an example, they pose the following explanatory question: "What does the proportion of people who ever told a lie in their life reveal about the prevalence of lying?"

For our purposes, however, the actual prevalence of individual untoward research behaviors is not as important as the fact that so many scientists do appear to engage in at least some of them and/or have observed others doing

so. Also, how we classify these behaviors is not particularly important, although one way to draw the line of demarcation is between ignorance and willfulness. *Everyone*, for example, knows that fabricating data is fraudulent, but not bothering to report all of ones variables might simply be due to substandard training and mentoring. Or it might be cultural based on the science in question.

However, while ignorance may have served as a passable excuse for engaging in some detrimental practices in the past, it does nothing to mitigate their deleterious effects on science and should no longer be tolerated given the amount of warnings promulgated in the past decade or so. For it is worth repeating that the ultimate effect of QRPs is the potential *invalidation* of the majority of some entire empirical literatures in an unknown number of scientific disciplines. And that, going forward, is simply not acceptable.

Modeling the Effects of Four Common Questionable Research Practices

Fortunately, just as a meta-analysis can be (and has been) conducted on just about every scientific topic imaginable, so can just about anything be modeled. And that latter truism (or unsubstantiated exaggeration) leads to one of the most iconoclastic (and definitely one of my favorite) articles in psychology—and most certainly possessive of one of that discipline's most pejorative titles.

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph Simmons, Leif Nelson, and Uri Simonsohn (2011)

This seminal article presents us with two genres of simulations quite different from the ones discussed in the previous chapter, the second of which is actually quite entertaining in a masochistic sort of way. But, regardless of the effect it has upon us, it is unlike anything that I have personally encountered in a peer reviewed journal.

Both models address the previous contention that the practice of a few inflationary QRPs allows savvy investigators (who must value reaping the rewards provided by their discipline over preserving its integrity) to increase their chances of producing statistically significant results to the point that a p-value < 0.05 is *more* likely to occur than a p-value > 0.05.

Simulation 1: *How to subvert the already generous alpha level of .05 and continue the decade- old process of constructing a trivial (but entertaining) science*.

This simulation is the more conventional of the two. Here 15,000 random samples were drawn from a normal distribution to assess the impact of the following four QRPs: (a) choosing which of two correlated outcome variables to report (plus an average of the two), (b) not specifying sample sizes a priori but beginning with 20 observations per cell and adding 10 more observations if statistical significance is not yet obtained, (c) using three experimental conditions and choosing whether to drop one (which produced four alternate analytic approaches), and (d) employing a dichotomous variable and its interaction with the four combinations of analytic alternatives detailed in (c).

Letting the computer do the heavy lifting, the p-values actually obtained for the 15,000 samples were computed and contrasted to three titular alpha levels (< 0.1, <0.05, and < .01). For the most commonly used level of 0.05, the percentage of false-positive results resulting from the four chosen scenarios were (recalling that a 5% rate would be expected to occur by chance alone) as follows:

(a) Use of two moderately correlated outcome variables (9.5%),
(b) Addition of 10 more Ss (7.7%) when statistical significance in not found,
(c) Dropping a treatment group or using all three (12.6%), and
(d) Adding a dichotomous variable plus its interaction with the treatments as a covariate (11.7%).

Now, as bad as this seems, anyone disingenuous (or untrained) enough to use any one of these strategies is also quite likely to use more than one (or even some additional QRPs) so the authors also included the effects of some of the combinations of the four which produced false-positive estimates ranging from 14.4% to a whopping 60.7%.

It is worth repeating that the just listed percentages assumed an alpha level of .05, but the authors also provided the same information for alphas of .10 and .01 (see table 1 in the original article). As would be expected, the QRP effects are much higher for an alpha of .10 (which is comparable to a one-tailed alpha of .05) and considerably lower for .01. (For example, the 14.4% and 60.7% range estimated for multiple QRP sins reduces to 3.3% and 21.5%, respectively, for an alpha of .01). And, as suggested by Benjamin, Berger, Johannesson, et al. (2017); Greenwald, Gonzalez, Harris, and Guthrie (1996); and the simulations in Table 2.1, the deleterious effects of the individual QRPs and their combinations would be greatly reduced if the titular alpha level were to be decreased to .005. However, since psychology and the vast majority of other scientific disciplines (at least the social sciences) aren't likely to adopt an alpha of .005 anytime soon, the criterion of 0.05 was employed in Simmons, Nelson, and Simonsohn's other astonishing simulation.

Simulation 2: *Also, how to subvert the already generous titular alpha level of .05 and continue the process of constructing an irreproducible (but entertaining) science.*

This one must surely be one of the first modeling strategies of its kind. Two experiments were reported, the first apparently legitimate (if trivial since it was based on Daryl Bem's infamous study "proving" that future events can influence past events) using a soft, single item and 30 undergraduates (also a typical sample size for psychology experiments) who were randomized to listen to one of two songs: "Hot Potato" (a children's tune that the undergraduates would most likely remember from childhood as the experimental condition) versus a rather blah instrumental control tune ("Kalimba"). Note that the experimental children's song was perfectly and purposefully selected to create an immediate reactive response by asking the undergraduates if they felt older immediately after listening to it. And, sure enough, employing the age of the participants' father as a covariate (which basically made no sense and was not justified) the experimental group listening to "Hot Potato" reported that they had felt significantly older ($p = .033$) on a 5-point scale (the study outcome) than the group hearing the nonreactive control song.

For their second study the authors performed a "conceptual" replication of the one just described, but this time embedding all four of the

computer-modeled QRPs from Simulation 1. However, the authors first reported the study's design, analytic procedure, and results *without* mentioning any of these QRPs, which made it read like an abstract of a typically "successful" psychology publication:

> Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either "When I'm Sixty-Four" by The Beatles or "Kalimba." Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father's age. We used father's age to control for variation in baseline age across participants. An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted M = 20.1 years) rather than to "Kalimba." (Adjusted M = 21.5 years), p = .040). (p. 1360)

Then, they confessed their sins:

1. A second intervention (listening to the Beatles' "When I'm Sixty-Four") was employed in lieu of "Hot Potato" ("Hot Potato" did not reach statistical significance this time around as compared to the "Kalimba" control and hence was dropped from the analysis and not mentioned in the simulated report);
2. The participants' (a) father's age, (b) gender, (c) and the gender's interaction with the experimental conditions were employed as covariates (a plethora of other variables were also included and apparently auditioned as covariates, such as participants' mother's age, their political persuasion, an item about Canadian quarterbacks, and so forth);
3. Following the analytic process a new outcome variable (the participants' *adjusted* ages) was employed since the first study's self-reported variable (i.e., whether the respondents felt older) also did not reach statistical significance this time; and
4. The analytic process itself included post hoc interim analyses conducted after additional participants were run until statistical significance was achieved—at which time the exercise was terminated.

The Authors' (Simmons, Nelson, and Simonsohn) Suggestions for Improvement

These suggestions were presented via two headings ("Requirements for Authors" and "Guidelines for Reviewers"). Some of these are obvious given the second simulation and some, unfortunately, may not go far enough. First the author requirements (these are in italics and numbered as they appear on pages 1362–1363 of the original article):

1. *Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.* Any serious institutional review board (IRB) requires such a statement in proposals submitted to them along with a rationale for prematurely terminating a study or adding more participants to the original sample size justification if applicable. Submission of these documents should probably be required by journals prior to publication if the same information is not preregistered.
2. *Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.* This one is unclear since an *N* of 68 per group is necessary to produce adequate statistical power for a typical social science effect size of 0.50. Ironically, a number of authors (e.g., Bakker, van Dijk, & Wicherts, 2012) have lamented the fact that the typical power available for psychological experiments can be as low as 0.35, and (also ironically) 20 participants per cell doesn't even quite meet this low criterion for a two-group study.
3. *Authors must list all variables collected in a study.* This one is quite important and of course should be an integral part of the preregistration process for relatively simple experiments such as the ones described here. Large clinical random controlled trials (RCTs) (as well as databases used for correlational studies) often collect a large number of demographic, background, health, and even cost data, a simple list of which might run several pages in length. Thus perhaps this suggestion could be loosened a bit for some types of research. However, those used as covariates, blocking variables, subgroup analyses, and, of course, primary outcomes must be prespecified accordingly.

4. *Authors must report all experimental conditions, including failed manipulations.* The failure to include an extra intervention or comparison group should be considered censorable misconduct.
5. *If observations are eliminated, authors must also report what the statistical results are if those observations are included*. And, of course, a rationale for the inclusion-exclusion criteria and the treatment of outliers (with definitions) should be provided a priori.
6. *If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.* This is an excellent point and is seldom adhered to. Additionally, the actual covariate–outcome correlation should be reported (which is almost never done). Covariates always adjust the meanings of outcomes to a certain extent, so it is extremely important to ensure that the adjusted outcome conceptually remains a variable of interest. It is not immediately apparent, for example, exactly what adjusting "feeling older" in the first experiment or adjusting "participants' ages" based upon "fathers' ages" in the second experiment winds up producing.

The second list of guidelines is presented for peer reviewers. These will be supplemented in Chapter 9, which discusses publishing concerns.

1. *Reviewers should ensure that authors follow the requirements* [presumably set by the journal or professional guidelines]. This is especially important for preregistration of key elements of the experimental process, as illustrated in the authors' second simulation. And as the authors note: "If reviewers require authors to follow these requirements, they will" (p. 1363).
2. *Reviewers should be more tolerant of imperfections in results*. "Underpowered studies with perfect results are the ones that should invite extra scrutiny" (p. 1363).
3. *Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions*. It might even be suggested that arbitrary analytic decisions shouldn't be made in the first place.
4. *If justifications of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication*. With apologies to Drs. Simmons, Nelson, and Simonsohn, one might wonder if it makes sense to perform an "exact" self-replication of a study with design flaws or non-compelling analytic procedures.

> While the original effect did not replicate in this simulation, might not some flaws that produced a false-positive result in an original study (such as an obvious demand characteristic coupled with a lack of blinding) also produce a false-positive result in an exact replication? Which leads to an even more bizarre question.

Would the Two Studies Conducted by the Simmons Team Replicate?

Now, as previously mentioned, I am quite fond of these studies but we all like research results that reinforce our biases and predilections. So what about the reproducibility of these two studies? Well, their authors tell us that the first one didn't replicate when employing identical procedures and analyses. But what about the second study?

There's no way of knowing short of replicating it, but that's the wrong question anyway. A better one might be

> Certainly the contrived "Hot Potato" experiment is a tour de force illustration of how skillful manipulations of QRBs are *capable* of creating nonreplicable, false-positive results, but do these miscreants also produce comparable false-positive result in the *published* literature?

Said another way, the Simmons et al. study demonstrated the effects that QRPs *could* have on the artifactual achievement of statistically significant results. So while the study may not have demonstrated (a) the actual occurrence of artifactually significant results or (b) that their four QRPs actually *do* result in artifactual statistical significance in the published literature, surely it demonstrated their *potential* for doing so.

Of course, the authors' first simulation involving the likely effects of their four key QRPs provides strong evidence that such practices also have the potential to dramatically inflate the obtained p-value and hence produce false-positive results—evidence buttressed by the preceding survey results demonstrating the high prevalence of these and other QRPs in the actual conduct of scientific research.

But even more convincingly, a group of management investigators fortuitously picked up where the Simmons et al. study left off and demonstrated the

actual effects of QRPs on the production of statistically significant findings. In my personal opinion this particular study provides one of the best *empirical* documentations of the untoward effects of QRPs and their implicit relationship to both publication bias and false-positive results.

The investigators accomplished this impressive feat by longitudinally following a group of studies from their authors' dissertation to their subsequent publication in peer reviewed journals. And, as if this wasn't sufficiently impressive, the title of their article rivals the iconic entries of both Ioannidis's ("Why Most Published Research Findings Are False") and Simmons et al.'s ("False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant").

The Chrysalis Effect: How Ugly Initial Results Metamorphosize into Beautiful Articles

Ernest O'Boyle, Jr., George Banks, and
Erik Gonzalez-Mule, E. (2014)

Without going into excruciating detail on the study methods, basically, the authors identified management-related dissertations registered between 2000 and 2012 that possessed a formal hypothesis and were subsequently published. The task wasn't as simple as it sounds since the published articles often had different titles and/or failed to mention that they were based on dissertation research (the latter being mandated by published ethics codes of the Academy of Management and the American Psychological Association).

Thankfully the investigators persevered and were able to identify 142 dissertations where there was "overwhelming" evidence that the studies had been subsequently published in a refereed journal. (The average time to publication was 3.29 years.) Altogether (i.e., in both dissertations and journal articles), there were 2,311 hypotheses, 1,978 of which were tested in the dissertations and 978 in the paired articles.

Overall differences between dissertation and journal article results showed that of the 1,978 hypotheses contained in the dissertations, 889 (44.9%) were statistically significant while, of the 978 hypotheses tested in the publications, 645 (65.9%) achieved that status. Or, in the authors' conceptualization of their results, "Our primary finding is that from

dissertation to journal article, the ratio of supported to unsupported hypotheses more than doubled (0.82 to 1.00 versus 1.94 to 1.00)" (p. 376).

Another way to view these results is to consider only the 645 hypothesis tests which were common to both dissertations and journal articles. Here, 56 of the 242 (20.6%) negative hypothesis tests in the dissertations somehow changed into positive findings in the published articles, while only 17 of the 323 (4.6%) positive dissertation findings were changed to negative ones. That in turn reflects a greater than four-fold negative to positive change as compared to a positive to negative one.

As for results due to QRPs, perhaps the most creative aspect of this seminal study involved estimating the effects of individual QRPs on the bottom-line inferential changes occurring over time for the 142 paired study versions. (Perhaps not coincidentally the five QRPs in this study basically overlap the four modeled in the Simmons et al. study, which, in all but one case, overlapped the preceding John et al. survey.)

QRP 1: *Deletion or addition of data after hypothesis tests*. Across the 142 projects, 14 (9.9%) added subjects (as evidenced by increases in sample size from dissertation to journal) and 29 (20.4%) dropped subjects. Overall both adding and deleting participants resulted in increased statistical significance over time (24.5% vs. 10.2%, respectively).

When broken down by adding versus deleting participants, 19% of the effects changed from negative to positive when the sample size was *increased*, while 8.9% (a two-fold reduction) changed in the opposite direction. (Note that this contrast was not statistically significant because of the relatively few studies that increased their sample size over time.) Among the studies that dropped subjects, there was a 2.5-fold difference favoring changes from non-significance to statistical significance as compared to positive to negative changes (28.1% vs. 11.1%, respectively).

QRP 2: Altering the data after hypothesis testing. This potential QRP was assessed in 77 studies in which the sample size did not change over time. (There were 22 cases in which it was not possible to determine whether data were altered.) The authors rationale was that "those studies that added or deleted data have a logical (*but not necessarily appropriate* [emphasis added]) reason why descriptive statistics would change from dissertations to their matched journal publications" (p. 386). Of these 77 studies, 25 (32.5%) showed changes in the means, standard deviations, or interrelations of the included variables, which represented 47 nonsignificant hypothesis tests and 63 statistically significant ones. Following

publication, 16 (34%) of the negative studies became positive and 0% changed from positive to negative.

QRP 3: *Selective deletion or addition of variables*. Deleting the same 22 studies as in QRP 2 (i.e., for which data alteration couldn't be ascertained) left 120 pairs (i.e., 142 – 22). Of these, 90 included instances "where not all of the variables included in the dissertation appeared in the publication and 63 (52.5%) instances where not all of the variables found in the article were reported in the dissertation." (There were 59 studies which both added and dropped variables.) In the dissertations, there were 84 negative tests and 136 positive ones. Adding variables to the published studies resulted in a change from negative to positive of 29.8% as compared to an 8.1% change in the negative direction—a three-fold migration favoring negative to positive change.

QRP 4: *Reversing the direction or reframing hypotheses to support data*. The authors note that this artifact doesn't necessarily include changing a hypothesis from "the intervention will be efficacious" to "it won't work." Instead it might involve adding a covariate (recall the iconoclastic Simmons et al. simulations) or slightly changing a predicted three-way interaction effect that might actually be significant as hypothesized but not in the expected direction. Here, the good news is that only eight studies representing 22 hypothesis tests were guilty of substantively reframing the original hypothesis. But of course "bad" news usually follows good news in this arena so, in this case, the bad news is that none (0%) of these 22 dissertation hypotheses was originally statistically significant while 17 (77.3%) of the p-values "somehow" changed from $p > 0.05$ to $p < 0.05$ when published.

QRP 5: *Post hoc dropping or adding of hypotheses*. Of the 142 paired studies, 126 (87.7%) either dropped or added a hypothesis, with 80 doing both. This translates to (a) 1,333 dropped hypotheses of which 516 (38.7%) were statistically significant as opposed to (b) 333 added hypotheses of which 233 (70.0%) were statistically significant. In other words, the new hypotheses were almost twice as likely to be positive as the ones that were dropped from the dissertation.

Qualifiers: The authors quite transparently describe a number of alternative explanations to some of their findings. For example, they note that

> between the dissertation defense and journal publication, it is possible, even likely, that mistakes were identified and corrected, outliers removed, new

> analytic techniques employed, and so on that would be classified as questionable by our criteria [i.e., labeled as QRPs in their study] but were nevertheless wholly appropriate to that particular project. That being said, these changes consistently coincided with increases in statistical significance and increases in the ratio of supported to unsupported hypotheses, and on this basis, we conclude that the preponderance of QRPs are engaged in for non-ideal [a nice euphuism] reasons. (p. 392)
>
> Other possible weaknesses identified by the authors in their data included the possibilities that some of the discrepancies noted between dissertation and journal publications might (a) have been mandated by journal editors or peer reviewers based on space limitations or (b) unique to the academic doctoral process (e.g., "forced" on the doctoral students by one or more committee members.)

However, it should be remembered that both the O'Boyle et al. and the Simmons et al. designs do not necessarily lead to definitive *causal* conclusions. With that said, I personally consider these studies to be extremely creative and both their data (actual or simulated) and the conclusions based on them quite persuasive, especially when considered in the context of the other previously presented observational and modeling studies coupled with actual replication results that will be presented shortly.

Perhaps, then, the primary contribution of this and the previous chapter's simulations resides in the facts that

1. The disciplinary-accepted or prespecified alpha level for any given study is probably almost *never* the actual alpha level that winds up being tested at study's end—unless, of course, a study is properly designed, conducted, and analyzed;
2. Many, many investigators—while complying with *some* important and obvious bias-reducing strategies—still conduct studies that are improperly designed, conducted, and analyzed (hence biased in other ways).

So although we may never be capable of ascertaining the true false-positive rate of any discipline (or even the actual correct p-value emanating from any imperfectly designed, conducted, and/or analyzed study), we do know that the percentage of false-positive results for most studies employing an alpha

level of 0.05 will be considerably above the 5% figure suggested in cell b of Table 2.1. And assuming the validity of the modeled false-positive rate for the presence of one or more of the four contraindicated practices employed in the Simmons et al. 2011 paper, the false-positive rate would be expected to mutate from the theoretical 5% level to between 7.7% and (a downright alarming) 60.7%.

But while this minimal estimate of an alpha increase of 2.7% (7.7% −5%) may not appear to be a particularly alarming figure, it is important to remember that this translates into tens of thousands of false-positive results. (However, it should also be noted that correlations among outcome variables greater than the 0.50 value employed in this simulation would result in a greater increase in the veridical alpha level.) And, even more discouraging, the projections for the untoward effects of both the four unitary QRPs and their combinations may well be underestimates. And to make matters worse still, the total menu of QRPs soon to be enumerated ensures many more options for producing fallacious statistically significant findings than those modeled in the Simmons et al. paper and validated in the later O'Boyle et al. study.

So perhaps this is a good time to introduce of few more QRPs. But perhaps we should first differentiate between investigator-driven QRPs and inane institutional policies (IIPs) which, like questionable investigator behavior, are also capable of contributing to the prevalence of irreproducible results.

A Partial List of Inane Institutional Scientific Policies and Questionable Research Practices

Inane institutional scientific policies (IISPs) are not directly under the personal control of individual investigators, but this does not imply that they cannot be changed over time by individual scientists through their own advocacy, professional behaviors, or collectively via group pressures. QRPs, on the other hand, are almost exclusively under individual investigators' personal control, although adversely influenced by IISPs and inadequate scientific mentorship.

So first consider the following list of the more common IISP culprits:

1. *Publication bias,* which has already been discussed in detail and is partially due to institutional behaviors involving journal editors, funders,

peer reviewers, publishers, the press, and the public, in addition to individual investigator behaviors. (So this one, like some that follow, constitutes combination QRP-IISP issues.) Researchers often bemoan the fact that journal editors and peer reviewers make negative studies so difficult to publish, but who, after all, are these nefarious and short-sighted miscreants? Obviously the vast majority are researchers themselves since they typically serve in these publishing and funding capacities. It is therefore incumbent upon these individuals to not discriminate against well-conducted nonsignificant studies and to so lobby the institutions for which they work.

2. A *flawed peer review system* that encourages publication bias, ignores certain QRPs, does not always enforce journal guidelines, and sometimes engages in cronyism. But again, who are these peer reviewers? The answer is that almost everyone reading this is (or will be) a peer reviewer at some point in their career. And some, heaven forbid, may even become journal editors which will provide them with an even greater opportunity to influence attitudes and practices among both their peer reviewers and their publishers. (Suggestions for reforming the peer review process will be discussed in some detail in Chapter 9.)
3. *Insufficient scientific mentoring and acculturation of new or prospective investigators*. This one is tricky because senior mentors have important experiential advantages but some are completely "set in their ways," resistant to change, and may not even be aware of many of the issues discussed in this book. However, one doesn't have to be long of tooth to adopt an informal mentoring role and guide new or prospective researchers toward the conduct of methodologically sound research.
4. *A concomitant lack of substantive disciplinary and methodological knowledge on the part of many of these insufficiently mentored investigators*. Some of the onus here lies with these individuals to supplement their own education via the many online or print sources available. However, institutions also bear a very real responsibility for providing ongoing educational opportunities for their new faculty researchers—as well as inculcating the need for self-education, which is freely and conveniently available online.
5. *Institutional fiscal priorities* resulting in untoward pressure to publish and attract external research funds. These issues are largely outside any single individual's personal control but someone must at least attempt

to educate the perpetrators thereof to change their policies—perhaps via organized group efforts.

6. Related to this is the academic administration's seeming adoption of the corporate model of perpetual expansion (physical and financial) resulting in evaluating department heads based on the amount of external grant funds their faculties manage to garner. Such pressures can force senior scientists to spend too much of their time pursuing funding opportunities at the expense of actually engaging in scientific activities or adequately supervising their staff who do. And many of the latter in turn understand that *they* will be evaluated on the number of publications their work generates, which leads us back to publication bias and the multiple geneses of the reproducibility crisis.
7. *The institutionalization of too many disciplines possessing no useful or truly unique knowledge base* and thereby ensuring the conduct of repetitive, obvious, and trivial research. (In the spirit of transparency, this is one of my personally held, idiosyncratic opinions.) One option for individuals stuck in such pseudoscientific professions is to seek opportunities to work with research teams in other more propitious arenas. Alternately (or additionally) they can try mightily to discover what might be a useful and/or parsimonious theory to guide research in their field, which in turn might eventually lead to a scientifically and societally useful knowledge base.
8. And related to Number 7 is the institutional practice of never abandoning (and seldom downsizing) one of these disciplines while forever creating additional ones. Or recognizing when even some mature, previously productive disciplines have exhausted their supply of "low hanging fruit" and hence may be incapable of advancing beyond it. This, of course, encourages trivial, repetitive, and obvious studies over actual discoveries as well as possibly increasing the pressure to produce exciting, counterintuitive (hence often false positive) findings. The only cure for this state of affairs is for investigators to spend less time conducting tautological studies and spend more time searching for more propitious avenues of inquiry. It is worth noting, however, that publishing obviously trivial studies whose positive effect are already known *should* result in statistical significance and not contribute to a discipline's false positive rate.

9. And related to *both* of the previous IISPs is the reluctance of funding agencies to grant awards to speculative or risky proposals. There was even once an adage at the National Institutes of Health (NIH) to the effect that the agency seldom funds a study to which the result isn't already known. But of course the NIH is not a stand-alone organism nor do its employees unilaterally decide what will be funded. Scientists are the ones who know what is already known, and they have the most input in judging what is innovative, what is not, what should be funded, and what shouldn't be.
10. *Using publication and citation counts as institutional requirements for promotion, tenure, or salary increases*. Both practices fuel the compulsion to publish as much, as often, and keyed to what investigators believe will result in the most citations as humanly possible. We all employ numeric goals in our personal life such as exercise, weight loss, and wealth (or the lack thereof), but excessive publication rates may actually decrease the probability of making a meaningful scientific contribution and almost certainly increases publication bias. One study (Ioannidis, Klavans, & Boyack, 2018) reported that, between 2000 and 2016, 9,000 individuals published one paper every 5 days. True, the majority of these were published in high-energy and particle physics (86%) where the number of co-authors sometimes exceeded a thousand, but papers in other disciplines with 100 co-authors were not uncommon. (Ironically the lead author of this paper [John Ioannidis, who I obviously admire] has published more than a thousand papers himself in which he was either first or last author. But let's give him a pass here.)

And Now a Partial List of Individual Questionable Research Practices

It should be noted that some of these QRPs are discipline- or genre-specific but the majority are applicable (perhaps with a bit of translation) to most types of empirical research. There is also some unavoidable interdependence among the list's entries (i.e., some share superordinate methodological components) as well as possessing similar redundancies with a number of the QRPs previously mentioned. But with these caveats and disclaimers dutifully disclosed, hopefully the following annotated list represents a

reasonably comprehensive directory of the most relevant QRPs to scientific irreproducibility:

1. *The use of soft, reactive, imprecise, self-reported, and easily manipulated outcome variables that are often chosen, created, or honed by investigators to differentially fit their interventions*. This one is especially endemic to those social science investigators who have the luxury of choosing or constructing their own outcomes and tailoring them to better match (or be more reactive to) one experimental group than the other. From a social science perspective, however, if an outcome variable has no social or scientific significance (such as how old participants feel), then the experiment itself will most likely also have no significance. But it will most likely add to the growing reservoir of false-positive results.
2. *Failure to control for potential experimenter and participant expectancy effects*. In some disciplines this may be the most virulent QRP of all. Naturally, double-blinded randomized designs involving sensible control/comparison groups are crucial in psychology and medicine, given demand and placebo effects, respectively, but, as will be demonstrated in Chapter 5, they are equally important in the physical sciences as well. As is the necessity of blinding investigators and research assistants to group membership in animal studies or in any research that employs variables scored by humans or that require human interpretation. (The randomization of genetically identical rodents and then blinding research assistants to group membership is a bit more complicated and labor intensive in practice than it may appear. Also untoward effects can be quite subtle and even counterintuitive, such as male laboratory rats responding differentially to male research assistants.)
3. *Failure to report study glitches and weaknesses*. It is a rare study in which no glitches occur during its commission. It is true, as Simmons et al. suggest, that some reviewers punish investigators for revealing imperfections in an experiment, but it has been my experience (as both a journal editor-in-chief and investigator) that many reviewers appreciate (and perhaps reward) transparency in a research report. But more importantly, hiding a serious glitch in the conduct of a study may result in a false-positive finding that of course can't be replicated.
4. *Selective reporting of result*s. This has probably been illustrated and discussed in sufficient detail in the Simmons et al. paper in which

selected outcomes, covariates, and experimental conditions were deleted and not reported in either the procedure or result sections. Another facet of this QRP involves "the misreporting of true effect sizes in published studies . . . that occurs when researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results" (Head, Holman, Lanfear, et al., 2015, p. 1). Suffice it to say that all of these practices are substantive contributors to the prevalence of false-positive results.

5. *Failure to adjust p-values based on the use of multiple outcomes, subgroup analyses, secondary analyses, multiple "looks" at the data prior to analysis, or similar practices resulting in artifactual statistical significance*. This is an especially egregious problem in studies involving large longitudinal databases containing huge numbers of variables which can easily yield thousands of potential associations there among. (Using nutritional epidemiology as an example, the European Prospective Investigation into Cancer and the Nutrition, Nurses' Health Study has resulted in more than 1,000 articles each [Ioannidis, 2018].) I personally have no idea what a reasonable adjusted p-value should be in such instances, although obtaining one close to 0.05 will obviously be completely irrelevant. (Perhaps somewhere in the neighborhood [but a bit more liberal] to the titular alpha levels adopted by the genomic and particle physics fields, which will be discussed later.)
6. *Sloppy statistical analyses and erroneous results in the reporting of p-values*. Errors such as these proliferate across the sciences, as illustrated by David Vaux's (2012) article in *Nature* (pejoratively titled "Know When Your Numbers Are Significant"), in which he takes biological journals and investigators to task for simple errors and sometimes absurd practices such as employing complex statistical procedures involving *N*s of 1 or 2. As another very basic example, Michal Krawczyk (2008), using a dataset of more than 135,000 p-values found (among other things) that 8% of them appeared to be inconsistent with the statistics upon which they were based (e.g., t or F) and, à propos of the next QRP, that authors appear "to round the p- values down more eagerly than up." Perhaps more disturbing, Bakker and Wicherts (2011), in an examination of 281 articles, found "that around 18% of statistical results in the psychological literature are incorrectly reported . . . and around 15% of the articles contained at least one statistical conclusion

that proved, upon recalculation, to be incorrect; that is, recalculation rendered the previously significant result insignificant, or vice versa" (p. 666). And it should come as no surprise by now that said errors were most often in line with researchers' expectations, hence an example of confirmation bias (Nickerson, 1998).

7. *Procedurally lowering p-values that are close to, but not quite ≤ 0.05*. This might include suspicious (and almost always unreported) machinations such as (a) searching for covariates or alternate statistical procedures or (b) deleting a participant or two who appears to be an outlier in order to whittle down a p of, say, 0.07 a couple of notches. Several investigators in several disciplines (e.g., Masicampo & Lalande, 2012; Gerber, Malhotra, Dowling, & Doherty, 2010; Ridley, Kolm, Freckelton, & Gage, 2007) have noted a large discrepancy between the proportions of p-values found just below 0.05 (e.g., 0.025 to 0.049) as opposed to those just above it (vs. 0.051 to 0.075).
8. *Insufficient attention to statistical power issues* (e.g., conducting experiments with too few participants). While most past methodology textbooks have emphasized the deleterious effects of low power on the production of negative studies, as previously discussed, insufficient power has equally (or greater) unfortunate effects on the production of false-positive findings. The primary mechanism of action of low power involves the increased likelihood of producing unusually large effect sizes by chance, which in turn are more likely to be published than studies producing more realistic results. And, as always, this QRP is magnified when coupled with others such as repeatedly analyzing results with the goal of stopping them as soon as an effect size large enough emerges to produce statistical significance. (Adhering to the prespecified sample size as determined by an appropriate power analysis would completely avoid this artifact.)
9. *Gaming the power analysis process*, such as by hypothesizing an unrealistically large effect size or not upwardly adjusting the required sample size for specialized designs such as those involving hierarchical or nested components.
10. *Artifactually sculpting (or selecting) experimental or control procedures to produce statistical significance* during the design process which might include:
    a. *Selecting tautological controls*, thereby guaranteeing positive results if the experiment is conducted properly. The best examples of this

involve comparisons between interventions which have a recognized mechanism of action (e.g., sufficient instruction delivered at an appropriate developmental level in education or a medical intervention known to elicit a placebo effect) versus "instruction or treatment as usual." (It could be argued that this is not a QRP if the resulting positive effect is not interpreted as evidence of efficacy, but, in the present author's experience, it almost always is [as an example, see Bausell & O'Connell, 2009]).

b. *Increasing the fit between the intervention group and the outcome variable or, conversely, decreasing the control–outcome match*. (The first, semi-legitimate experiment described in the Simmons et al. paper provides a good example of this although the study was also fatally underpowered from a reproducibility perspective.) While speculative on the present author's part, one wonders if psychology undergraduates or Amazon Mechanical Turk participants familiar with computer-administered experiments couldn't surmise that they had been assigned the experimental group when a song ("Hot Potato" which they had heard in their childhood) was interrupted by asking them how old they felt. Or if their randomly assigned counterparts couldn't guess that they were in the control group when a blah instrumental song was similarly interrupted. (Relatedly, Chandler, Mueller, and Paolacci [2014] found [a] that investigators tend to underestimate the degree to which Amazon Turk workers participate across multiple related experiments and [b] that they "overzealously" exclude research participants based on the quality of their work. Thirty-three percent of investigators employing crowdsourcing participants appear to adopt this latter approach, thereby potentially committing another QRP [i.e., see Number 11].)

11. *Post hoc deletion of participants or animals for subjective reasons*. As an extreme example, in one of my previous positions I once witnessed an alternative medicine researcher proudly explain his criterion for deciding which observations were legitimate and which were not in his animal studies. (The latter's lab specialized in reputably demonstrating the pain-relieving efficacy of acupuncture resulting from tiny needles being inserted into tiny rat legs and comparing the results to a placebo.) His criterion for deciding which animals to delete was proudly explained as "sacrificing the non-acupuncture responding

animals" with an accompanying smile while drawing his forefinger across his neck. (No, I am not making this up nor do I drink while writing. At least not at this moment.)

12. *Improper handling and reporting of missing data*. Brief-duration experiments normally do not have problematically high dropout rates, but interventions whose effects must be studied over time can suffer significant attrition. Preferred options for compensating for missing data vary from discipline to discipline and include regression-based imputation of missing values (available in most widely used statistical packages) and intent-to-treat. (The latter tending to be more conservative than the various kinds of imputation and certainly the analysis of complete data only.) Naturally, the preregistration of protocols for such studies should describe the specific procedures planned, and the final analyses should comply with the original protocol, preferably presenting the results for both the compensatory and unvarnished data. Most funding and regulatory agencies for clinical trials require the prespecification of one or more of these options in their grant proposals, as do some IRBs for their submissions. Of course it should come as no surprise that one set of investigators (Melander, Ahlqvist-Rastad, Meijer, & Beermann, 2003) found that 24% of standalone studies neglected to include their preregistered intent-to-treat analysis in the final analysis of a cohort of antidepressant drug efficacy experiments—presumably because such analyses produce more conservative (i.e., less positive) results than their counterparts. (Because of the high financial stakes involved, positive published pharmaceutical research results tend to be greatly facilitated by the judicious use of QRPs [see Turner, Matthews, Linardatos, et al., 2008, for an especially egregious example]).
13. *Adding participants to a pilot study in the presence of a promising trend, thereby making the pilot data part of the final study*. Hopefully self-explanatory, although this is another facet of performing interim analyses until a desired p-value is obtained.
14. *Abandoning a study prior to completion based on the realization that statistical significance is highly unlikely to occur (or perhaps even that the comparison group is outperforming the intervention)*. The mechanism by which this behavior inflates the obtained p-value may not be immediately apparent, but abandoning an ongoing experiment (i.e., not a pilot study) before its completion based on (a) the perceived

impotence of the intervention, (b) the insensitivity of the outcome variable, or (c) a control group that might be performing above expectations allows an investigator to conserve resources and immediately initiate another study until one is found that is sufficiently promising. Continually conducting such studies until statistical significance is achieved ultimately increases the prevalence of false, non-replicable positive results in a scientific literature while possibly encouraging triviality at the same time.

15. *Changing hypotheses based on the results obtained.* Several facets of this QRP have already been discussed, such as switching primary outcomes and deleting experimental conditions, but there are many other permutations such as (a) obtaining an unexpected result and presenting it as the original hypothesis or (b) reporting a secondary finding as a planned discovery. (All of which fit under the concepts of *HARKing* [for Hypothesizing After the Results are Known, Kerr, 1991] or *p-hacking* [Head et al., 2015], which basically encompasses a menu of strategies in which "researchers collect or select data or statistical analyses until nonsignificant results become significant.") As an additional example, sometimes a plethora of information is collected from participants for multiple reasons, and occasionally one unexpectedly turns out to be influenced by the intervention or related to another variable. Reporting such a result (or writing another article based on it) without explicitly stating that said finding resulted from an exploratory analysis constitutes a QRP in its own right. Not to mention contributing to publication bias and the prevalence of false-positive results.
16. *An overly slavish adherence to a theory or worldview*. *Confirmation bias*, a tendency to search for evidence in support of one's hypothesis and ignore or rationalize anything that opposes, it is subsumable under this QRP. A more extreme manifestation is the previously mentioned animal lab investigator's literal termination of rodents when they failed to respond to his acupuncture intervention. But also, perhaps more commonly, individuals who are completely intellectually committed to a specific theory are sometimes capable of actually seeing phenomenon that isn't there (or failing to see disconfirmatory evidence that is present). The history of science is replete with examples such as craniology (Gould, 1981), cold fusion (Taubes, 1993), and a number of other pathologies which will be discussed in

Chapter 5. Adequate controls and effective blinding procedures are both simple and absolutely necessary strategies for preventing this very troublesome (and irritating) QRP.

17. *Failure to adhere to professional association research standards and journal publishing "requirements,"* such as the preregistration of statistical approaches, primary hypotheses, and primary endpoints before conducting studies. Again, as Simmons and colleagues state: "If reviewers require authors to follow these requirements, they will" (p. 1363).
18. *Failure to provide adequate supervision of research staff.* Most experimental procedures (even something as straightforward as the randomization of participants to conditions or the strict adherence to a standardized script) can easily be subverted by less than conscientious (or eager to please) research staff, so a certain amount of supervision (e.g., via irregular spot checks) of research staff is required.
19. *Outright fraud,* of which there are myriad, well-publicized, and infamous examples, with perhaps the most egregious genre being data fabrication such as (a) painting patches on mice with permanent markers to mimic skin grafts (Hixson, 1976), (b) Cyril Burt making a splendid career out of pretending to administer IQ tests to phantom twins separated at birth to "prove" the dominance of "nature over nurture" (Wade, 1976), or (c) Yoshitaka Fujii's epic publication of 172 fraudulent articles (Stroebe, Postmes, & Spears, 2012). While data fabrications such as these are often dismissed as a significant cause of scientific irreproducibility because of their approximately 2% self-reported incidence (Fanelli, 2009), even this probable underestimate is problematic when one considers the millions of entries in published scientific databases.
20. *Fishing, data dredging, data torturing* (Mills, 1993)*, and data mining.* All of which are used to describe practices designed to reduce a p-value below the 0.05 (aka p-hacking) threshold by analyzing large numbers of variables in search of statistically significant relationships to report—but somehow forgetting to mention the process by which these findings were obtained.
21. *The combined effects of multiple QRPs*, which greatly compounds the likelihood of false-positive effects since (a) a number of these practices are independent of one another and therefore their effects on false-positive results are cumulative and (b) individuals who knowingly

commit one of these practices will undoubtedly be inclined to combine it with others when expedient.

22. *Failing to preregistering studies and ensure their accessible to readers*. It is difficult to overemphasize the importance of preregistering study protocols since this simple strategy would avoid many of the QRPs listed here *if* preregistrations are routinely compared to the final research reports during the peer review process. Or, barring that, they are routinely compared by bloggers or via other social media outlets.
23. *Failure to adhere to the genre of established experimental design standards discussed in classic research methods books and the myriad sets of research guidelines discussed later.* Common sense examples include (a) randomization of participants (which should entail following a strict, computer-generated procedure accompanied by steps to blind experimenters, participants, and principal investigators); (b) the avoidance of experimental confounds, the assurance of reliability, and the validity of measuring instruments, taking Herculean steps (if necessary) to avoid attrition; and (c) a plethora of others, all of which should be common knowledge to anyone who has taken a research methods course. However, far and away the most important of these (with the possible exception of random assignment) is the blinding of experimenters (including animal and preclinical researchers), research assistants, and participants (including everyone who comes in contact with them) with respect to group membership and study hypotheses/purposes. Of course this is not possible in some genres of research, as when having research assistants count handwashing episodes in public lavatories (Munger & Harris, 1989), employing confederates in obedience studies (Milgram, 1963), or comparing the effects of actual knee surgery to placebo surgery (Moseley, O'Malley, Petersen, et al., 2002), but in most research scenarios blinding can and must be successfully instituted.

It may be that imperfect blinding of participants and research staff (at least those who come in contact with participants) may be among the most virulent QRPs in experiments in which investigators tend to *sculpt experimental or control* procedures to produce statistical significance (QRPs Numbers 2 and 10) and/or are themselves metaphorically blinded, given the degree to which they are wedded to their theory or word view (QRP Number 16).

If this is true, it follows that investigators should (and should be required to) employ blinding *checks* to ascertain if the procedures they put (or failed to put) into place were effective. Unfortunately a considerable amount of evidence exists that this seemingly obvious procedure is seldom employed. Much of this evidence comes from individual trials, such as the classic embarrassment in which 311 NIH employees were randomly assigned to take either a placebo or ascorbic acid capsule three times a day for 9 months to ascertain the effectiveness of vitamin C for the treatment and prevention of the common cold. Unfortunately the investigators failed to construct a placebo that matched the acidic taste of the intervention, and a blinding check revealed that many of the NIH participants broke said blind attempt by tasting the capsules (Karlowski, Chalmers, Frenkel, et al., 1975).

Unfortunately a number of methodologists have uncovered substandard blinding efforts (most notably failures to evaluate their effectiveness) in a number of different types of studies. Examples include

1. Fergusson, Glass, Waring, and Shapiro (2004) found that only 8% of 191 general medical and psychiatric trials reported the success of blinding;
2. A larger study (Baethge, Assall, & Baldessarini, 2013) found even worse results, with only 2.5% of 2,467 schizophrenia and affective disorders RCTs reported assessing participant, rater, or clinician blinding; and, not to be outdone,
3. Hróbjartsson, Forfang, Haahr, and colleagues (2007) found an ever lower rate (2%) of blinding assessments for a sample of 1,599 interdisciplinary blinded RCTs.

However, every scientist residing outside of a cave knows that participants, research assistants, and clinicians *should* be blinded (at least when feasible, since studies involving surgery or acupuncture cannot blind the individuals administering the treatments). Unfortunately compliance with this knowledge is a bit short of perfect.

Colagiuri and Benedetti (2010), for example, quite succinctly sum up the importance of universal blinding checks in a carefully crafted criticism of the otherwise excellent CONSORT 2010 Statement's updated guidelines for randomized trials which inexplicably deleted a provision to check blinding and downgraded it to a recommendation. The Colagiuri and Benedetti team explained the rationale for their criticism (via a *British Medical Journal* Rapid Response article):

> Testing for blinding is the *only* [emphasis added] valid way to determine whether a trial is blind. Trialists conducing RCTs should, therefore, report on the success of blinding. In situations where blinding is successful, trialists and readers can be confident that guesses about treatment allocation have not biased the trial's outcome. In situations where blinding fails, trialists and readers will have to evaluate whether or not bias may have influenced the trial's outcomes. Importantly, however, in the second situation, while trialists are unable to label their trial as blind, the failure of blinding should not be taken as definitive evidence that bias occurred. Instead, trialists should provide a rationale as to why the test of blinding was unsuccessful and a statement on whether or not they consider the differences between treatment arms to be valid. (2010, p. 340)

While these authors' comments were directed at medical researchers, one wonders just how successful double-blinding is in the average low-powered social science experiment—especially given the high prevalence of positive results in these latter literatures. Why not, therefore, take the time to administer a one-item blinding check to participants after their completion of the experimental task by simply asking to which treatment group they believed they had been assigned? This is especially important in psychology since Amazon Mechanical Turk participants and psychology students are probably more experienced and savvy in gleaning the purpose of a study than investigators realize.

Comparing participants' guesses regarding assignment to their actual assignment would be an excellent mechanism for evaluating the success of whatever blinding strategy was implemented. And perhaps the very act of evaluating this strategy might induce investigators to be more careful about the design of their studies and their crafting of experimental conditions. (Especially if the blinding check was included in the preregistration.) So let's add the following, perhaps idiosyncratic, QRP to our burgeoning (but still incomplete) list:

24. *Failing to check and report experimental participants' knowledge (or guesses) regarding the treatments they received.* Again, this could be done by administering a single force-choice item (e.g., "To what condition [experimental or control] do you believe you were assigned?") at the end of an experiment and then correlating said answer not only to actual group assignment but also to ascertain if there was an

interaction between said guesses and actual assignment with respect to outcome scores. The answers would not necessarily be causally definitive one way or another, but a large portion of the participants correctly guessing their treatment assignment by study's end would be rather troublesome. And it would be even more troublesome if individuals in the control group who suspected they were in the intervention scored differentially higher or lower on the outcome variable than their control counterparts who correctly guessed their group assignment.

A Few "Not Quite QRPs" but Definitely Irritating Reporting Practices

These do not necessarily impact either publication bias or the prevalence of false-positive results, but they are, at the very least, disingenuous, irritating, and possibly becoming more prevalent.

1. *Downplaying or failing to mention study limitations* (Ioannidis, 2007);
2. *Hyping results* via such increasingly common descriptors as "robust, novel, innovative, and unprecedented" (Vinkers, Tijdink, & Otte, 2015); and
3. *"Spinning" study results*; Boutron, Dutton, Ravaud, and Altman (2010), for example, found that "a majority of 72 trials reporting non-statistically significant results had included *spin* type descriptors in both the abstract and the conclusion sections [which the authors found particularly problematic since many clinicians read only those two sections]. A number of these even attributed beneficial effects for the treatment being evaluated despite statistically *nonsignificant results*.

As a counter-example to these hyping and spinning practices, consider the announcement of James Watson and Francis Crick's (the former who definitely did not having a propensity for minimizing his accomplishments) introductory statement in the paper announcing the most heralded biological discovery of the twentieth century: "We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest" (1953, p. 737).

QRPs and Animal Studies

While there are many different types of laboratory experiments, those employing in vivo animals undoubtedly have the most in common with human experimentation and hence are most susceptible to the types of QRPs of concern to us here. There is, in fact, a surprising similarity in requisite behaviors and procedures required for producing valid, reproducible results in the two genres of research—at least with a bit of translation.

Certainly some investigative behaviors are unique to humans, such as the necessity of ensuring that participants can't guess their group membership or querying them thereupon (QRP Number 24). But, as with experiments employing human participants, a disquieting amount of literature exists chronicling the shortcomings of published in vivo animal studies. In fact it is probably safe to say that, historically, preclinical investigators have been among the leaders in failing to report their methods thoroughly and their adherence to recommended experimental practices.

This is especially problematic in animal studies that precede and inform randomized clinical human trials designed to test the efficacy of pharmaceutical therapies. Unfortunately, the track record for how well animal studies actually do inform human trials borders on the abysmal. It has been estimated, for example, that only about 11% of therapeutic agents that have *proved promising* in animal research and have been tested clinically are ultimately licensed (Kola & Landis, 2004). In addition, Contopoulos-Ioannidis, Ntzani, and Ioannidis (2003) found that even fewer (5%) of "high impact" basic science discoveries claiming clinical relevance are *ever* successfully translated into approved therapies within a decade.

Naturally, every preclinical study isn't expected to result in a positive clinical result, but these statistics relate to *positive* animal studies, not those that initially "failed." It is therefore highly probable that a significant number of the former reflect false-positive results due to many of the same QRPs that bedevil human experimentation.

This supposition is buttressed by Kilkenny, Parsons, Kadyszewski, and colleagues (2009) who, surveying a large sample of published animal studies, found that only 13% reported randomizing animals to treatments and only 14% apparently engaged in blinded data collection. This particular study, incidentally, apparently led to the Animal Research: Reporting of in Vivo Experiments (ARRIVE) guidelines (Kilkenny, Browne, Cuthill, et al., 2010) which is closely modeled on the CONSORT statement—yet another commonality between the two genres of research. And, like its

predecessor, ARRIVE also has its own checklist (http://www.nc3rs.org.uk/ARRIVEchecklist/) and has likewise been endorsed by an impressive number of journals.

So, as would be expected, the majority of the ARRIVE procedural reporting guidelines (e.g., involving how randomization or blinding was performed if instituted) are similar to their CONSORT counterparts although others are obviously unique to animal research. (For a somewhat more extensive list of methodological suggestions for this important genre of research, see Henderson, Kimmelman, Fergusson, et al., 2013.)

Unfortunately, while the ARRIVE initiative has been welcomed by animal researchers and a wide swath of preclinical journals, enforcement has been a recurring disappointment, as demonstrated in the following study title "Two Years Later: Journals Are Not Yet Enforcing the ARRIVE Guidelines on Reporting Standards for Pre-Clinical Animal Studies" (Baker, Lidster, Sottomayor, & Amor, 2014). The investigators, in examining a large number of such studies published in the journals *PLoS* and *Nature* (both of which officially endorsed the ARRIVE reporting guidelines) found that

1. The reporting of blinding "was similar to that in past surveys (20% in *PLoS* journals and 21% in *Nature* journals)," and
2. "Fewer than 10% of the relevant studies in either *Nature* or *PLoS* journals reported randomisation (10% in *PLoS* journals and 0% in *Nature* journals), and even fewer mentioned any power/sample size analysis (0% in *PLoS* journals and 7% in *Nature* journals)" (p. 3).

From one perspective, perhaps 2 years is not a great deal of time for comprehensive guidelines such as these to be implemented. But from a scientific perspective, this glass is neither half full nor half empty because behaviors such as blinding, randomization, and power analyses should not require guidelines in the 21st century. Rather their commission should be ironclad prerequisites for publication.

Whether the primary etiology of this disappointing state of affairs resides in journal policies, mentoring, or knowledge deficiencies is not known. However, since hopefully 99% of practicing scientists know that these three methodological procedures are absolute requirements for the production of valid experimental findings, the major onus probably involves journal involvement, such as the suggestion by Simmons, Nelson, and Simonsohn (2012) (based on their iconic 2011 article) that investigators affix a 21-word statement to their published experiments (i.e., "We report how we

determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study"). A simple innovation which inspired both the PsychDisclosure initiative (LeBel, Borsboom, Giner-Sorolla, et al., 2013) and the decision of the editor of the most prestigious journal in the field (*Psychological Science*) to *require* authors' disclosure via a brief checklist. (Checklists, incidentally have been shown to be important peer review aids and are actually associated with improved compliance with methodological guidelines [Han, Olonisakin, Pribis, et al., 2017]).

Next Up

The next chapter features a discussion of some especially egregious case studies graphically illustrating the causal link between QRPs and irreproducibility, the purpose of which is not to pile demented examples upon one another but rather to suggest that the QRP → irreproducibility chain will undoubtedly be more difficult to sever than we would all like.

References

Baethge, C., Assall, O. P., & Baldessarini, R. J. (2013). Systematic review of blinding assessment in randomized controlled trials in schizophrenia and affective disorders 2000–2010. *Psychotherapy and Psychosomatics, 82*, 152–160.

Baker, D., Lidster, K., Sottomayor, A., & Amor, S. (2014). Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biology*, *11*, e1001756.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research, 43*, 666–678.

Bausell, R. B., & O'Connell, N. E. (2009). Acupuncture research: Placebos by many other names. *Archives of Internal Medicine, 169*, 1812–1813.

Bedeian, A. G., Taylor, S. G., & Miller, A. N. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education*, *9*, 715–725.

Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2017). Redefine statistical significance. *Nature Human Behavior, 2*, 6–10.

Boutron, I., Dutton, S., Ravaud, P., & Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Journal of the American Medical Association*, *303*, 2058–2064.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaivete among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods, 46*, 112–130.

Colagiuri, B., & Benedetti, F. (2010). Testing for blinding is the only way to determine whether a trial is blind. *British Medical Journal, 340*, c332. https://www.bmj.com/rapid-response/2011/11/02/testing-blinding-only-way-determine-whether-trial-blind

Contopoulos-Ioannidis, D. G., Ntzani, E., & Ioannidis, J. P. (2003). Translation of highly promising basic science research into clinical applications. *American Journal of Medicine, 114*, 477–484.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*, e5738.

Fergusson, D., Glass, K. C., Waring, D., & Shapiro, S. (2004). Turning a blind eye: The success of blinding reported in a random sample of randomized, placebo controlled trials. *British Medical Journal, 328*, 432.

Fiedler, K., & Schwarz N. (2015). Questionable research practices revisited. *Social Psychological and Personality Science 7*, 45–52.

Gerber, A S., Malhotra, N., Dowling, C. M, & Doherty, D. (2010). Publication bias in two political behavior literatures. *American Politics Research, 38*, 591–613.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology, 33*, 175–183.

Han, S., Olonisakin, T. F., Pribis, J. P., et al. (2017). A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. *PLoS ONE*, *12*, e0183591.

Head, M. L., Holman, L., Lanfear, R., et al. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*, e1002106.

Henderson, V. C., Kimmelman, J., Fergusson, D., et al. (2013). Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. *PLoS Medicine, 10*, e1001489.

Hixson, J. R. (1976). *The patchwork mouse*. Boston: Anchor Press.

Hróbjartsson, A., Forfang, E., Haahr, M. T., et al. (2007). Blinded trials taken to the test: An analysis of randomized clinical trials that report tests for the success of blinding. *International Journal of Epidemiology, 36*, 654–663.

Ioannidis, J. P. A. (2007). Limitations are not properly acknowledged in the scientific literature. *Journal of Clinical Epidemiology, 60*, 324–329.

Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2018). Thousands of scientists publish a paper every five days, papers and trying to understand what the authors have done. *Nature, 561*, 167–169.

Ioannidis, J. P. A. (2018). The challenge of reforming nutritional epidemiologic research. *JAMA, 320*, 969–970.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23*, 524–532.

Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., et al. (1975). Ascorbic acid for the common cold: A prophylactic and therapeutic trial. *Journal of the American Medical Association, 231*, 1038–1042.

Kerr, N. L. (1991). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Kilkenny, C., Browne, W. J., Cuthill, I. C., et al. (2010). Improving bioscience research reporting: The ARRIVE Guidelines for reporting animal research. *PLoS Biology, 8*, e1000412.

Kilkenny, C., Parsons, N., Kadyszewski, E., et al. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE, 4*, e7824.

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery, 3*, 711–715.

Krawczyk, M. (2008). *Lies, Damned lies and statistics: The adverse incentive effects of the publication bias*. Working paper, University of Amsterdam. http://dare.uva.nl/record/302534

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., et al. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8*, 424–432.

Masicampo E. J., & Lalande D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology, 65*, 2271–2279.

Melander, H., Ahlqvist-Rastad, J., Meijer, G., & Beermann, B. (2003). Evidence b(i)ased medicine-selective reporting from studies sponsored by pharmaceutical industry: Review of studies in new drug applications. *British Medical Journal, 326*, 1171–1173.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–378.

Mills, J. L. (1993). Data torturing. *New England Journal of Medicine, 329*, 1196–1199.

Moseley, J. B., O'Malley, K., Petersen, N. J., et al. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *New England Journal of Medicine*, *347*, 82–89.

Munger, K., & Harris, S. J. (1989). Effects of an observer on hand washing in public restroom. *Perceptual and Motor Skills*, *69*, 733–735.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220.

O'Boyle, Jr., E. H., Banks, G. C., & Gonzalez-Mule, E. (2014). The Chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*, 376–399.

Ridley, J., Kolm, N., Freckelton, R. P., & Gage, M. J. G. (2007). An unexpected influence of widely used significance thresholds on the distribution of reported P-values. *Journal of Evolutionary Biology, 20*, 1082–1089.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. https://ssrn.com/abstract=2160588

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*, 670–688.

Taubes, G. (1993). *Bad science: The short life and weird times of cold fusion*. New York: Random House.

Turner, E. H., Matthews, A. M., Linardatos, E., et al. (2008) Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252–260.

Vaux D. (2012). Know when your numbers are significant. *Nature, 492*, 180–181.

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *British Medical Journal, 351*, h6467.

Wade, N. (1976). IQ and heredity: Suspicion of fraud beclouds classic experiment. *Science, 194*, 916–919.

Watson, J. D., & Crick, F. (1953). A structure for deoxyribose nucleic acid. *Nature, 171*, 737–738.

4

A Few Case Studies of QRP-Driven Irreproducible Results

The title of our first entry unambiguously describes the article's bottom-line conclusions as well as several of the other case studies presented in this chapter.

The Statistical Crisis in Science: Data-Dependent Analysis—A "Garden of Forking Paths"—Explains Why Many Statistically Significant Comparisons Don't Hold Up

Andrew Gelman and Eric Loken (2014)

The authors illustrate their point by briefly mentioning several psychological studies that either failed to replicate (e.g., women being more likely to wear pink or red at peak fertility [Beall & Tracy, 2013]) or that report unrealistically large effect sizes (e.g., women changing their voting preferences based on their ovulatory cycles [Durante, Rae, & Griskevicius, 2013]). While these and other examples are linked to specific statistical questionable research practices (QRPs) listed in the previous chapter, Gelman and Loken metaphorically articulate the problem in terms of taking a scientific journey in which there are many forks in the road, thereby necessitating the many decisions that must be made before arriving at the final destination.

Or, in their far more eloquent words:

> In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done *implicitly* [emphasis added]. The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientific common sense

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0005

> to formulate their hypotheses in a reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, this is strong evidence in favor of the hypothesis. (p. 464)

The authors go on to rather compassionately suggest how such practices can occur without the conscious awareness of their perpetrators.

> Working scientists are also keenly aware of the risks of data dredging, and they use confidence intervals and p-values as a tool to avoid getting fooled by noise. Unfortunately, a by-product of all this struggle and care is that when a statistically significant pattern does show up, it is natural to get excited and believe it. The very fact that scientists generally don't cheat, generally don't go fishing for statistical significance, makes them vulnerable to drawing strong conclusions when they encounter a pattern that is robust enough to cross the $p < 0.05$ threshold. (p. 464)

As for solutions, the authors suggest that, in the absence of preregistration, almost all conclusions will be data-driven rather than hypothesis-driven. And for observational disciplines employing large *unique* databases (hence impossible to replicate using different data), they suggest that *all of the relevant comparisons* in such databases be more fully analyzed while not concentrating only on statistically significant results.

But now let's begin our more detailed case studies by considering a hybrid psychological-genetic area of study involving one of the former's founding constructs (general intelligence or the much reviled [and reified] *g*). We won't dwell on whether this construct even exists or not, but for those doubters who want to consider the issue further, Jay Gould's *Mismeasure of Man* (1981) or Howard Gardner's (1983) *Frames of Mind: The Theory of Multiple Intelligences* are definitely recommended.

What makes the following study so relevant to reproducibility are (a) its demonstration of how a single QRP can give rise to an entire field of false-positive results while at the same time (b) demonstrating the power of the *replication process* to identify these fallacious results.

Most Reported Genetic Associations with General Intelligence Are Probably False Positives

Christopher Chabris, Benjamin Herbert,
Daniel Benjamin, et al. (2012)

In their introduction, these authors justify their study by arguing that "General cognitive ability, or *g,* is one of the most heritable behavioral traits" (p. 1315). (Apparently a large literature has indeed found statistically significant correlations between various single-nucleotide polymorphisms [SNPs; of which there are estimated to be ten million or so] and *g*.)

The present example represented a *replication* of 13 of the SNPs previously found to be related to *g* in an exhaustive review by Antony Payton (2009) spanning the years from 1995 to 2009; these SNPs happened to be located near 10 potentially propitious genes. The authors' replications employed three large, "well-characterized" longitudinal databases containing yoked information on at least 10 of these 13 SNPs: (a) the Wisconsin high school students and a randomly selected sibling ($N = 5,571$), (b) the initial and offspring cohorts of the Framingham Heart Study ($N = 1,759$), and (c) a sample of recently genotyped Swedish twins born between 1936 and 1958 ($N = 2,441$).

In all, this effort resulted in 32 regression analyses (controlling for variables such as age, gender, and cohort) performed on the relationship between the appropriate data on each SNP and IQ. Only one of these analyses reached statistical significance at the 0.04 level (a very low bar), and it was in the *opposite direction to that occurring in one of the three original studies*. Given the available statistical power of these replications and the number of tests computed, the authors estimate that at least 10 of these 32 associations should have been significant at the .05 level by chance alone.

Their conclusions, while addressed specifically to genetic social science researchers, is unfortunately relevant to a much broader scientific audience.

> Associations of candidate genes with psychological traits and other traits studied in the social sciences should be viewed as tentative until they

> have been replicated in multiple large samples. Failing to exercise such caution may hamper scientific progress by allowing for the proliferation of potentially false results, which may then influence the research agendas of scientists who do not realize that the associations they take as a starting point for their efforts may not be real. And the dissemination of false results to the public may lead to incorrect perceptions about the state of knowledge in the field, especially knowledge concerning genetic variants that have been described as "genes for" traits on the basis of unintentionally inflated estimates of effect size and statistical significance. (p. 1321)

A Follow-Up

Partially due to replication failures such as this one (but also due to technology and lowering costs of a superior alternative), candidate gene analyses involving alpha levels of .05 have largely disappeared from the current scientific literature. Now genome-wide association studies are in vogue, and the genetics research community has reduced the significance criterion by several orders of magnitude (i.e., $p \leq .0000005$). Obviously, as illustrated in the simulations discussed in Chapter 2, appropriate reductions in titular alpha levels based on sensible criteria will greatly reduce the prevalence of false-positive results, and perhaps, just perhaps, psychology will eventually follow suit and reduce its recommended alpha level for new "discoveries" to a more sensible level such as 0.005.

Unfortunately, psychological experiments would be considerably more difficult to conduct if the rules were changed in this way. And much of this science (and alas others as well) appears to be a game, as suggested by the Bakker, van Dijk, and Wicherts (2012) title ("The Rules of the Game Called Psychological Science") along with many of the previously discussed articles. But let's turn our attention to yet another hybrid psychological-physiological foray somewhere past the boundaries of either irreproducibility or inanity, this time with a somewhat less pejorative title.

Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition

Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler (2009)

We've all been regaled by functional magnetic resonance imaging (fMRI) studies breathlessly reporting that one region or another of the brain is associated with this or that attribute, emotion, or whatever. Thousands have been conducted, and 41 meta-analyses were located by the intrepid and tireless John Ioannidis who definitively documented an excess of statistically significant findings therein (2011).

However, the Vul et al. study went several steps further than is customary in research such as this. First, the authors examined not only statistical significance but the effect sizes produced, noting that many of the Pearson correlations coefficients (which serve as effect sizes in correlational research) between fMRI-measured brain activity and myriad other psychosocial constructs (e.g., emotion, personality, and "social cognition") were in excess of 0.80. (Correlation coefficients range from −1.00 to +1.00, with +/−1.00 representing a perfect correspondence between two variables and zero indicating no relationship whatever.) Then, given the "puzzling" (some would unkindly say "highly suspicious") size of these correlations, the team delved into the etiology of these astonishingly large (and decidedly suspect) values.

Suspect because veridical correlations (as opposed to those observed by chance, data analysis errors, or fraud) of this size are basically impossible in the social sciences because the reliability (i.e., stability or reliability) of these disciplines' measures typically fall *below* 0.80 (and the reliability of neuroimaging measures, regardless of the disciplines involved, usually falls a bit short of 0.70). Reliability, as the authors of this study note, places an algebraic upper limit on how high even a perfect correlation (i.e., 1.00) between two variables can be achieved via the following very simple formula:

Formula 4.1: *The maximum correlation possible between two variables given their reliabilities*

$$\begin{aligned}&\text{Corrected Perfect Correlation}\\&= \sqrt{\begin{matrix}\text{Measure 1}\left(\text{Reliability of Psychological Measures}\right)\\ \times \text{Reliability of fMRI Mesurements)}\end{matrix}}\\&= \sqrt{.80\left(\text{Psychological Measure}\right) \times .70\left(\text{FMRI Measures}\right)}\\&= \sqrt{.56} = 0.75\end{aligned}$$

So, to paraphrase Shakespeare, perhaps something is rotten in brain imaging research? Alas, let's all hope that it's only brain imaging research's olfactory output that is so affected.

To get to the bottom of what the authors charitably described as "puzzling," a literature search was conducted to identify fMRI studies involving correlations between the amount of deoxygenated hemoglobin in the blood (called the BOLD signal, which is basically a measure of blood flow) and psychosocial constructs as measured by self-reported questionnaires. The search "resulted in 55 articles, with *274 significant correlations* [emphasis added] between BOLD signal and a trait measure" (p. 276), which if nothing else is a preconfirmation of Ioannidis's previously mentioned 2011 analysis concerning the extent of publication bias in fMRI research in general.

The next step in teasing out the etiology of this phenomenon involved contacting the authors of the 55 studies to obtain more information on how they performed their correlational analyses. (Details on the fMRI data points were not provided in the journal publications.) Incredibly, at least some information was received from 53 of the 55 articles, close to a record response for such requests and is probably indicative of the authors' confidence regarding the validity of their results.

Now to understand the etiology of these "puzzling" correlation coefficients, a bit of background is needed on the social science stampede into brain imaging as well as the analytics of how increased blood flow is measured in fMRI studies in general.

First, Vul and colleagues marveled at the eagerness with which the social sciences (primarily psychology) had jumped into the neuroimaging fad only a few years prior to the present paper, as witnessed by

1. The creation of at least two new journals (*Social Neuroscience* and *Social Cognitive and Affective Neuroscience*),
2. The announcement of a major funding initiative in the area by the National Institute of Mental Health in 2007, and (and seemingly most impressive to Vul and his co-authors),
3. "The number of papers from this area that have appeared in such prominent journals as *Science*, *Nature*, and *Nature Neuroscience*" (p. 274). (The first two of these journals have been accused of historically exhibiting a penchant for publishing sensationalist, "man bites dog" studies while ignoring the likelihood of their being reproducible.)

Second, three such studies were briefly described by the authors (two published in *Science* and one in a journal called *NeuroImage*) that associated brain activity in various areas of the brain with self-reported psychological scales while

1. Playing a game that induced social rejection (Eisenberger, Lieberman, & Williams, 2003),
2. Completing an empathy-related manipulation (Singer, Seymour, O'Doherty, et al., 2004), or
3. Listening to angry versus neutral speech (Sander, Grandjean, Pourtois, et al., 2005).

Incredibly, the average correlation for the three studies was 0.77. But, as explained earlier, this is statistically impossible given the amount of error (i.e., 1– reliability) present in both the psychological predictors and the fMRI scans. So even if the true (i.e., error-free) relationship between blood flow in the studied regions and the psychological scales was *perfect* (i.e., 1.00), the maximum numerically possible correlation among these variables in our less than perfect scientific world, would be *less* than 0.77. Or, said another way, the results obtained were basically statistically impossible since practically no social science scales are measured with sufficient precision to support such high correlations.

The authors also kindly and succinctly provided a scientific explanation (as opposed to the psychometric one just tendered) for why perfect correlations in studies such as these are almost impossible.

> First, it is far-fetched to suppose that only one brain area influences any behavioral trait. Second, even if the neural underpinnings of a trait were confined to one particular region, it would seem to require an *extraordinarily favorable set of coincidences* [emphasis added] for the BOLD signal (basically a blood flow measure) assessed in one particular stimulus or task contrast to capture all functions relevant to the behavioral trait, which, after all, reflects the organization of complex neural circuitry residing in that brain area. (p. 276)

And finally, to complete this brief tour of fMRI research, the authors provide a very clear description of how brain imagining "works," which I will attempt to abstract without botching it too badly.

A functional scanning image is comprised of multiple blood flow/oxygenation signals from roughly cube-shaped regions of the brain called voxels (*volumetric pixels*, which may be as small as 1 mm^3 or as large as 125 mm^3). The number of voxels in any given image typically ranges from 40,000 to 500,000 of these tiny three-dimensional pieces of the brain, and the blood flow within each of these can be correlated with *any* other data collected on the individual (in this case, psychosocial questionnaires involving self-reports).

Each voxel can then be analyzed separately with any variable of interest available on (or administered to) the individuals scanned—normally 20 or fewer participants are employed (sometimes considerably fewer) given the expense and machine time required. As mentioned in the present case, these dependent variables are psychosocial measures of perhaps questionable validity but which can reflect *anything*. The intervention can also encompass a wide range of manipulations, such as contrasting behavioral scenarios or gaming exercises.

Thus we have a situation in which there are potentially hundreds of thousands of correlations that *could* be run between each voxel "score" and a single independent variable (e.g., a digital game structured to elicit an emotional response of some sort or even listening to "Hot Potato" vs. "Kalimba," although unfortunately, to my knowledge, that particular study has yet to be conducted via the use of fMRI). Naturally, reporting thousands of correlation coefficients would take up a good deal of journal space so groups of voxels are selected in almost any manner investigators choose in order to "simplify" matters. And therein resides the solution to our investigators' "puzzlement" because the following

mind-bending strategy was employed by a *majority* of the investigators of the 53 responding authors:

> First, the investigator computes a separate correlation of the behavioral measure of interest with each of the voxels (fig. 4 in the original article). Then, he or she selects those voxels that exhibited a sufficiently high correlation (by passing a statistical threshold; fig. 4b). Finally, an ostensible measure of the "true" correlation is aggregated from the [subset of] voxels that showed high correlations (e.g., by taking the mean of the voxels over the threshold). With enough voxels, such a biased analysis is *guaranteed* [emphasis added] to produce high correlations even if none are truly present [i.e., by chance alone]. Moreover, this analysis will produce visually pleasing scatter grams (e.g., fig. 4c) that will provide (quite meaningless) reassurance to the viewer that s/he is looking at a result that is solid, is "not driven by outliers," and so on. . . . This approach amounts to selecting one or more voxels based on a functional analysis and then reporting the results of the same analysis and functional data from just the selected voxels. This analysis distorts the results by selecting noise that exhibits the effect being searched for, and any measures obtained from a non-independent analysis are biased and untrustworthy. (p. 279)

As problematic as this strategy is of reporting only the mean correlation involved in a subset of voxels highly correlated with the psychosocial intervention, 38% of the respondents actually "reported the correlation of the *peak* [emphasis added] voxel (the voxel with the highest observed correlation)" (p. 281).

The authors conclude their truly astonishing paper by suggesting alternative strategies for reducing the analytic biases resulting from this genre of research, including (a) ensuring that whoever chooses the voxels of interest be *blinded* to the voxel–behavioral measure correlations and (b) not to "peek" at the behavioral results while analyzing the fMRI output.

Another Perspective on the fMRI Studies

Undoubtedly the title of another article ("Voodoo Correlations Are Everywhere: Not Only in Neuroscience") represents a nod to Robert Park's

well-known book titled *Voodoo Science: The Road from Foolishness to Fraud* (2000). The article (Fiedler, 2011) is mentioned here because it also makes mention of the Vul et al. article but places it within the broader framework regarding the effects of idiosyncratic sampling decisions and their influence on the reproducibility of a much wider swath of scientific enquiry.

While Professor Fiedler acknowledges that "sampling" procedures such as the ones just discussed involving brain scanning studies are unusually egregious, he argues that psychological research as a whole is plagued with selective sampling strategies specifically designed to produce not only p-values ≤ .05, but what Roger Giner-Sorolla (2012) terms *aesthetically pleasing* results to go along with those propitious p-values. But let's allow Professor Fiedler to speak for himself in the study abstract by referring to the "voodoo correlations" just discussed.

> Closer inspection reveals that this problem [the voxel correlations] is only a special symptom of a broader methodological problem that characterizes *all paradigmatic research* [emphasis added] not just neuroscience. *Researchers not only select voxels to inflate effect size, they also select stimuli, task settings, favorable boundary conditions, dependent variables and independent variables, treatment levels, moderators, mediators, and multiple parameter settings in such a way that empirical phenomena become maximally visible and stable* [emphasis added again because this long sentence encapsulates such an important point]. In general, paradigms can be understood as conventional setups for producing idealized, inflated effects. Although the feasibility of representative designs is restricted, a viable remedy lies in a reorientation of paradigmatic research from the visibility of strong effect sizes to genuine validity and scientific scrutiny. (p. 163)

Fiedler's methodological language can be a bit idiosyncratic at times, such as his use of the term "metacognitive myopia" to characterize "a tendency in sophisticated researchers, who only see the data but overlook the sampling filters behind, [that] may be symptomatic of an industrious period of empirical progress, accompanied by a lack of interest in methodology and logic of science" (p. 167). However, he minces no words in getting his basic message across via statements such as this:

> Every step of experimental design and scholarly publication is biased toward strong and impressive findings, starting with the selection of a

> research question; the planning of a design; the selection of stimuli, variables and tasks; the decision to stop and write up an article; the success to publish it; its revision before publication; and the community's inclination to read, cite and adopt the results. (p. 167)

And while this statement may seem uncomfortably radical and unnecessarily pejorative, a good bet would be that the vast majority of the reproducibility experts cited in this book would agree with it. As they might also with the statement that "as authors or reviewers, we have all witnessed studies not published because [the experimental effect] was too small, but hardly any manuscript was rejected because the treatment needed for a given outcome was too strong" (p. 166).

But is it really fair to characterize the results of fMRI studies such as these as "voodoo"? Robert Park certainly would, but it isn't important how we classify QRP-laden research such as this. What is important is that we recognize the virulence of QRPs to subvert the entire scientific process and somehow find a way to agree upon a means to reduce the flood of false-positive research being produced daily. For it is all types of empirical studies and many entire scientific literatures, not just psychology or experimental research, that have ignited the spark that has grown into a full-blown initiative designed to illuminate and ultimately deflate what is a full-blown crisis. But since it may appear that psychology is being unduly picked on here, let's visit yet another discipline that is at least its equal in the production of false-positive results.

Another Irreproducibility Poster Child: Epidemiology

Major components of this science involve (a) case control studies, in which individuals with a disease are compared to those without the disease, and (b) secondary analyses of large databases to tease out risk factors and causes of diseases. (Excluded here are the roots of the discipline's name, tracking potential epidemics and devising strategies to prevent or slow their progress—obviously a vital societal activity upon which we all depend.)

Also, while case control studies have their shortcomings, it the secondary analysis wing of the discipline that concerns us here. The fodder for these analyses is mostly comprised of surveys, longitudinal studies (e.g., the Framingham Heart Study), and other large databases (often constructed for other purposes but fortuitously tending to be composed of large numbers of

variables with the potential of providing hundreds of secondary analyses and hence publications).

The most serious problems with such analyses include already discussed QRPs such as (a) data involving multiple risk factors and multiple conditions which permit huge fishing expeditions with no adjustments to the alpha level, (b) multiple confounding variables which, when (or if) identifiable can only be partially controlled by statistical machinations, and (c) reliance on self-reported data, which in turn relies on faulty memories and under- or overreporting biases.

These and other problems are discussed in a very readable article titled "Epidemiology Faces Its Limits," written more than two decades ago (Taubes, 1995) but which still has important scientific reproducibility implications today.

Taubes begins his article by referencing conflicting evidence emanating from analyses designed to identify cancer risk factors—beginning with those garnering significant press coverage during the year previous to the publication of his article (i.e., 1994).

1. Residential radon exposure caused lung cancer, and yet another study that found it did not.
2. DDT exposure was not associated with breast cancer, which conflicted with the findings of previous, smaller, positive studies.
3. Electromagnetic fields caused brain cancer, which conflicted with a previous study.

These examples are then followed by a plethora of others, a sample of which is included in the following sentence:

> Over the years, such studies have come up with a "mind-numbing array of potential disease-causing agents, from hair dyes (lymphomas, myelomas, and leukemia) to coffee (pancreatic cancer and heart disease) to oral contraceptives and other hormone treatments (virtually every disorder known to woman). (p. 164)

Of course we now know the etiology and partial "cures" for preventing (or at least lowering the incidence and assuaging the impact of) false-positive results (key among them in the case of epidemiology, and in specific analyses of large databases in particular, being lowering the titular alpha level

for such analyses to at *least* 0.005). However, while none of the plethora of epidemiologists Taubes interviewed in 1995 considered this option, several did recognize the related solution of considering only large relative risks (epidemiology's effect size of choice) including a charmingly nontechnical summarization of the field articulated by Michael Thun (the then-director of analytic epidemiology for the American Cancer Society): "With epidemiology you can tell a little [i.e., effect size] from a big thing. What's very hard to do is to tell a little thing from nothing" (p. 164).

And, to be fair, among the welter of false-positive results, we do owe the discipline for a few crucially important and exhaustively replicated relationships such as smoking and lung cancer, overexposure to sunlight and skin cancer, and the ill effects of obesity. (Coincidentally or not, all were discovered considerably before 1995. And definitely not coincidentally, all qualified as "big" effect sizes rather than "little" ones.)

Unfortunately while this article was written more than two decades ago, things don't appear to have improved all that much. And fortunately, one of my virtual mentors rescued me from the necessity of predicting what (if any) changes are likely to occur over the next 20 years.

Still, Taubes suggests (under the heading "What to Believe?") that the best answer to this question is to believe only strong correlations between diseases and risk factors which possess "a highly pausible biological mechanism."

Other steps are suggested by Stanley Young and Alan Karr (2011), including preregistration of analysis plans and (for large-scale observational studies) the use of a data cleaning team separate from the data analyst. These two epidemiologists also gift us with a liturgy of counterintuitive results that have failed to replicate, such as:

> Coffee causes pancreatic cancer. Type A personality causes heart attacks. Trans-fat is a killer. Women who eat breakfast cereal give birth to more boys. [Note that women do not contribute the necessary Y chromosome for producing male babies.] (p. 116)

Their next volley involved the presentation of a set of 12 observational studies involving 52 significant relationships that were subsequently subjected to randomized clinical trials. None of the 52 associations replicated, although some resulted in statistical results in the *opposite* direction.

Next, an observational study conducted by the US Centers for Disease Control (CDC) was described which correlated assayed urine samples for

275 chemicals with 32 medical outcomes and found, among other things, that bisphenol A was associated with cardiovascular disease, diabetes, and a marker for liver problems. However, there was a slight problem with this finding, but let's allow the authors to speak for themselves:

> There are 275 × 32 = 8800 potential endpoints for analysis. Using simple linear regression for covariate adjustment, there are approximately 1000 potential models, including or not including each demographic variable [there were 10]. Altogether the search space is about 9 million models and endpoints. The authors remain convinced that their claim is valid. (p. 120)

And finally a special epidemiological favorite of mine, involving as it does the secondary analysis of the data emanating from (of all things) one of the largest and most important randomized experiments in the history of education evaluating the learning effects of small versus large class sizes (Word, Johnston, Bain, et al., 1990; Mosteller, 1999).

The Effect of Small Class Sizes on Mortality Through Age 29 Years: Evidence from a Multicenter Randomized Controlled Trial

Peter Muennig, Gretchen Johnson, and Elizabeth Wilde (2011)

These authors, in excellent epidemiological tradition, obtained the original data emanating from what is often referred to as the Tennessee Class-Size study and linked it to the National Death Index records to identify the former's relationship to the latter (i.e., the relationship between students who had been randomly assigned to large vs. small classrooms and their subsequent deaths between 1985 and the end of 2007). The authors' rationale for the study was succinctly stated as:

> A large number of nonexperimental studies have demonstrated that improved cognition and educational attainment are associated with large health benefits in adulthood [seven citations were provided]. However, the short-term health effects of different schooling policies are largely

> unknown, and long-term effects have never been evaluated using a randomized trial. (p. 1468)

And there is a good reason why a randomized trial has never investigated the long-term health implications of an instructional strategy. Why should anyone bother?

The authors thus imply that correlating an extraneous, non-hypothesized variable occurring subsequent to a randomized study constitutes a causal relationship between the intervention and the dependent variable (death in this case)—which of course is absurd, but absurdity in some disciplines seldom precludes publication.

What the authors found was that, through age 29, the students randomized to the small class size group experienced statistically significantly higher mortality rates than those randomized to regular size classes. (The original experiment also consisted of class sizes with and without an assigned teacher's aide, but fortunately the presence of an aide didn't turn out to be lethal.) Therefore in the authors' words:

> Between 1985 and 2007, there were 42 deaths among the 3,024 Project STAR participants who attended small classes, 45 deaths among the 4,508 participants who attended regular classes, and 59 deaths among the 4,249 participants who attended regular classes with an aide. (p. 1468)

Interestingly, the authors never discuss the possibility that this might be a chance finding or that the relationship might be non-causal in nature or that the entire study was an obvious fishing expedition. Instead they came up with the following mechanism of action:

> It is tempting to speculate that the additional attention children received in their smaller classes—and possibly in the classes with teacher's aide—helped them to become more outgoing and affirmed their intellectual curiosity. . . . However, this will occasionally have negative outcomes. Poisonings, drugs, drinking and driving, and firearms account for a greater degree of exploration (e.g., poisonings in childhood) and extroversion (e.g., social drug use in adolescence). However, this hypothesis remains highly speculative [*Do you think?*]. (p. 1473)

> It is also tempting to speculate that this finding might have run counter to the investigators' original expectations, given their above-quoted rationale for the study and the fact that the class size study in question resulted in significant learning gains (i.e., "improved cognition and educational attainment are associated with large health benefits in adulthood)" (p. 1468)

Epistemologically, the results of this study can probably be explained in terms of the existence of a number of hidden QRPs, such as not reporting multiple analyses involving multiple "outcome" variables or *harking* (a previously mentioned acronym [but soon to be discussed in more detail] invented by Norbert Kerr [1991] for *h*ypothesizing *a*fter the *r*esults are *k*nown).

Suffice it to say that much of epidemiology involves locating and often (as in this case) combining datasets, bleeding them of all the potential bivariate relationship possible, and then throwing in multiple covariates until something (anything) surfaces with a p-value ≤ 0.05. And persistent investigators will find such a p-value, no matter how many analyses or different datasets are required because, sooner or later, a statistically significant relationship *will* occur. And with a little creativity and a lot of chutzpah, said finding can be accompanied by some sort of mechanism of action. (In this case there is absolutely nothing in the entire educational literature [or within the much more rarified confines of common sense] to suggest that an innocuous behavioral intervention such as smaller class sizes, which have been proved to increase learning, are also lethal.)

Unfortunately, epidemiological practice doesn't appear to have changed a great deal between Taubes's 1995 article and this one published more than a decade later—or as of this writing more than two decades later. Just recently we have found that alcohol consumption in moderation is protective against heart disease, or it isn't, or that red wine is but white wine isn't, or that all alcohol is bad for you regardless of amount. Fish oil prevents heart attacks or it doesn't; this vitamin or supplement prevents this or that disease or it doesn't; antioxidants are the best drugs since antibiotics, or they aren't; vigorous exercise is beneficial or it isn't if someone sits at their desk too long and doesn't walk enough; ad nauseam.

So what are the solutions to this tragedy of inanities? The same as for psychology and most other similarly afflicted disciplines, some which have been

discussed, some which will be discussed, and some with no plausible mechanism of action and are so methodologically sloppy that they can be perfunctorily dismissed as not falling under the rubric of legitimate science.

This latter category happens to constitute the subject matter of our next chapter, which deals with a topic sometimes referred to as "pathological" science. It is certainly several steps down the scientific ladder from anything we've discussed so far, but it must be considered since it is actually a key component of our story.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554.

Beall, A. T., & Tracy, J. L. (2013). Women are more likely to wear red or pink at peak fertility. *Psychological Science, 24*, 1837–1841.

Chabris, C., Herbert, B., Benjamin, D., et al. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science, 23*, 1314–1323.

Durante, K., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science, 24*, 1007–1016.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science, 302*, 290–292.

Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist, 102*, 460–465.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*, 562–571.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Ioannidis, J. P. (2011). Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry, 68*, 773–780.

Kerr, N. L. (1991). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*, 113–127.

Muennig, P., Johnson, G., & Wilde, E. T. (2011). The effect of small class sizes on mortality through age 29 years: Evidence from a multicenter randomized controlled trial. *American Journal of Epidemiology, 173*, 1468–1474.

Park, R. (2000). *Voodoo science: The road from foolishness to fraud*. New York: Oxford University Press.

Payton, A. (2009). The impact of genetic research on our understanding of normal cognitive ageing: 1995 to 2009. *Neuropsychology Review, 19*, 451–477.

Sander, D., Grandjean, D., Pourtois, G., et al. (2005). Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. *NeuroImage, 28*, 848–858.

Singer, T., Seymour, B., O'Doherty, J., et al. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*, 1157–1162.

Taubes, G. (1995). Epidemiology facts its limits. *Science, 269*, 164–169.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

Word, E., Johnston, J., Bain, H. P., et al. (1990). *Student/teacher achievement ratio (STAR): Tennessee's K–3 class size study: Final summary report 1985–1990.* Nashville: Tennessee Department of Education.

Young, S. S., & Karr, A. (2011). Deming, data and observational studies: A process out of control and needing fixing. *Significance, 9*, 122–126.

5

The Return of Pathological Science Accompanied by a Pinch of Replication

In 1953, Irving Langmuir (a Nobel laureate in chemistry) gave what must have been one of the most entertaining colloquia in history (*Colloquim on Pathological Science*). Fortunately, it was recorded and later transcribed for posterity by R. N. Hall (http://galileo.phys.virginia.edu/~rjh2j/misc/Langmuir.pdf) for our edification.

In the talk, Professor Langmuir discussed a number of blind alleys taken over the years by scientists who failed to understand the importance of employing appropriate experimental *controls*, thereby allowing them to see effects "that just weren't there." Today we lump these failures under the methodological category of *blinding*—the failure of which, as mentioned previously, is one of the most pernicious and common of the questionable research practices (QRPs).

Several examples of counterintuitive physical science discoveries ascribable to this failing were presented by Professor Langmuir, such as n-rays in 1903 and mitogenic rays a couple of decades later. Both findings elicited huge excitement in the scientific community because they were unexplainable by any known physical theory of their times (i.e., had no plausible mechanism of action) and consequently generated hundreds (today this would be thousands) of scientific publications. Both ray genres gradually fell out of favor following the realization that their existence could only be observed by zealots and could not be *replicated* or *reproduced* when experimenter expectations were effectively eliminated via blinding controls.

However, lest we look back at these miscues too condescendingly, many of us can probably vaguely recall the cold fusion debacle of a couple of decades ago in Utah, which suffered a similar natural history: huge excitement, considerable skepticism, a few early positive replications, followed by many more failures to replicate, and capped by everyone in science except the lunatic fringe soon moving on. (Note that here the term "failure to replicate" does not refer to a replication of an original study not being undertaken but

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0006

instead to a divergent result and conclusion being reached from a performed replication.)

While the cold fusion episode will be discussed in a bit more detail later, an additional psychological example proffered by Langmuir is probably even more relevant to our story here, given the chemist's par excellence personal investigation into research purporting to prove the existence of extrasensory perception (ESP) conducted by the Duke psychologist Joseph Banks Rhine (1934).

At the time (1934), Langmuir was attending a meeting of the American Chemical Society at Duke University and requested a meeting with Rhine, who quite enthusiastically agreed due to Langmuir's scientific eminence. Interestingly, the visitor transparently revealed his agenda at the beginning of the meeting by explaining his opinion about "the characteristics of those things that aren't so" and that he believed these applied to Rhine's findings.

Rhine laughed and said "I wish you'd publish that. I'd love to have you publish it. That would stir up an awful lot of interest. I'd have more graduate students. We ought to have more graduate students. This thing is so important that we should have more people realize its importance. This should be one of the biggest departments in the university."

Fortunately for Duke, the athletic department was eventually awarded that status but let's return to our story. Basically, Rhine revealed that he was investigating both *clairvoyance*, in which an experimental participant was asked to guess the identity of facedown cards, and *telepathy*, in which the participant was required to read the mind of someone behind a screen who knew the identity of each card. Both reside in the extrasensory perception realm, as did Daryl Bem's precognition studies that had such a large impact on the reproducibility initiative.

As designed, these experiments were quite easy to conduct and the results should have been quite straightforward since chance occurrence was exactly five (20%) correct guesses with the 25-card deck that Rhine employed. His experimental procedures also appeared fine as well (at least for that era), so, short of fraud or some completely unanticipated artifact (such as occurred with the horse Clever Hans who exhibited remarkable arithmetic talents [https://en.wikipedia.org/wiki/Clever_Hans]), there was no obvious way the results could have been biased. (*Spoiler alert*: all scientific results *can* be biased, purposefully or accidentally.)

After conducting thousands of trials (Langmuir estimated hundreds of thousands, and he was probably good with numbers), Rhine found that

his participants correctly guessed the identity of the hidden cards 28% of the time. On the surface this may not sound earth-shattering but given the number of experiments conducted, the probability associated with these thousands upon thousands of trials would undoubtedly be equivalent to a randomly chosen individual winning the Mega Millions lottery twice in succession.

Needless to say this result met with a bit of skepticism on Langmuir's part, and it wasn't greatly assuaged when Rhine mentioned that he had (a) filled several filing cabinets with the results of experiments that had produced only chance results or lower and (b) taken the precaution of sealing each file and placing a code number on the outside because he "Didn't trust anybody to know that code. Nobody!"

When Langmuir impoliticly (after all he was a guest even if an enthusiastically welcomed one) expressed some incredulity at Rhine's ignoring such a mountain of negative evidence locked away on the theory that his distractors had deliberately guessed incorrectly just to "spite" him, Rhine was not in the least nonplussed. After a bit more probing on Langmuir's part, Rhine did amend his reason for not at least mentioning these negative results in his book (1934) on the topic to the fact that he hadn't had time to digest their significance and, furthermore, didn't want to mislead the public.

Naturally, Rhine's work has been replicated, but his results have not (Hines, 2003). As an interesting aside, in preparation for his meeting at Duke, Langmuir even "commissioned" his own replication by convincing his nephew, an employee of the Atomic Energy Commission at the time, to recruit some of his friends to spend several of their evenings attempting to replicate Rhine's experiments. At first the group became quite excited because their results (28% or 7 correct guesses out of 25) almost perfectly reflected those of Rhine's, but soon thereafter the results regressed down to chance (i.e., 5 correct guesses out of 25 cards).

Langmuir concluded his fascinating talk as follows (remember this is a transcription of a poorly recorded, informal lecture):

> The characteristics of [the examples he discussed], they have things in common. These are cases where there is no dishonesty involved but where people are tricked into false results by a lack of understanding about what human beings can do to themselves in the way of being led astray by subjective effects, wishful thinking or threshold interactions. These are examples of *pathological science* [emphasis added]. These are things that attracted a

> great deal of attention. Usually hundreds of papers have been published upon them. Sometimes they have lasted for fifteen or twenty years and then they gradually die away. (p. 13 of Hall's transcript of Langmuir's *Colloquim on Pathological Science* (1953).

Langmuir may have been a bit too kind in his conclusion of no dishonesty being involved since Rhine reported in his second book that the coded file cabinets originally designated as containing his enemies' purposely incorrect guesses had somehow morphed into an average of seven (28%) correct guesses—thereby exactly replicating his initially reported (and unfiled) results.

But in the final analysis it doesn't matter whether we label substandard scientific procedures as fraud or ignorance or stupidity or QRPs. What is important is that we must avoid allowing them to produce a crisis of confidence in the scientific process itself, not to mention impeding scientific progress in an era in which society increasingly depends on it.

So let's transition now to one of the parents of the modern reproducibility awakening. "Transition," because there is a centuries-old tradition of this genre of experimentation dating back at least to Benjamin Franklin's dismissal of Franz Mesmer's (another early parapsychologist) discovery of animal magnetism at the behest of King of France (Kaptchuk, 1999).

The Odd Case of Daryl Bem

In 2011, an apparently well-respected psychologist and *psi* devotee (*psi* is a "branch" of ESP involving precognition) named Daryl Bem published in the prestigious *Journal of Personality and Social Psychology* a series of nine experiments conducted over a 10-year period and entitled "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." The author's premise was that some individuals (or at least some Cornell undergraduates) could not only predict future events a la Joseph Banks Rhine work but, going one step farther, that the direction of causation is not limited to past events influencing future ones, but can travel from the future to the past.

Rather than describing all nine experiments, let's allow Bem to describe the basic methodology of his last two presented experiments, which were very similar in nature and, of course, positive. (Authors of multiexperiment studies often save what they consider to be the most definitive studies for last

and sometimes even present a negative study first to emphasize later, more positive findings). His abbreviated description of the eighth experiment's objective follows:

> Inspired by the White Queen's (a character in Lewis Carroll's *Through the Looking Glass—And What Alice Found There*) claim, the current experiment tested the hypothesis that memory can "work both ways" by testing whether rehearsing a set of words makes them easier to recall— even if the rehearsal takes place after the recall test is given. Participants were first shown a set of words and given a free recall test of those words. They were then given a set of practice exercises on a randomly selected subset of those words. The psi hypothesis was that the practice exercises would retroactively facilitate the recall of those words, and, hence, participants would recall more of the to-be-practiced words than the unpracticed words. (p. 419)

And who could argue with such a venerable theoretician as the White Queen? So to shorten the story a bit, naturally the hypothesis was supported:

> The results show that *practicing a set of words* after the recall test *does, in fact, reach back in time* [emphasis added] to facilitate the recall of those words. (p. 419)

After all, what else could it have been but "reaching back into time?" Perhaps William of Occam (unquestionably my most demanding virtual mentor) might have come up with some variant of his hair-brained parsimony principle, but the good man's work is *really* outdated so let's not go there.

To be fair, Professor Bem does provide a few other theoretical mechanisms of action related to quantum mechanics emanating from "conversations" taking place at an "interdisciplinary conference of physicists and psi researchers sponsored by the American Association for the Advancement of Science" (Bem, 2011, p. 423). (Perhaps some of which were centered around one such application advanced by homeopathy advocates to explain how water's *memory* of a substance—the latter dissolved therein but then completely removed—can still be there and be palliative even though the original substance elicited anti-palliative symptoms.)

Perhaps due to the unusual nature of the experiments, or perhaps due to Bem's repeated quoting of Carroll's White Queen (e.g., "memory works both ways" or "It's a poor sort of memory that only works backwards"), some psychologists initially thought the article might be a parody. But most didn't

since the study of psi has a venerable history within psychological research, and one survey (Wagner & Monnet, 1979) found that almost two-thirds of academic psychologists believed that psi was at least possible. (Although in the discipline's defense, this belief was actually lower for psychologists than for other college professors.)

However, it soon became apparent that the article was presented in all seriousness and consequently garnered considerable attention both in the professional and public press—perhaps because (a) Bem was an academically respectable psychological researcher (how "academic respectability" is bestowed is not clear) housed in a respectable Ivy League university (Cornell) and (b) the experiments' methodology seemed to adequately adhere to permissible psychological research practice at the time. (Permissible at least in the pre-reproducibility crisis era but not so much now, since, as mentioned by another Nobel Laureate, "the times they are [or may be] a changing.'")

And therein lay the difficulties of simply ignoring the article since the methodological quality of the study mirrored that of many other published studies and few psychologists believed that Professor Bem would have been untruthful about his experimental methods, much less have fabricated his data. So, playing by the rules of the game at the time and presenting considerable substantiating data, it could be argued that Bem's methods were marginally adequate. The problem was that by 2011 (the publication date of the article in question), (a) the rules of the game were actually beginning to change and (b) everyone beyond the lunatic fringe of the discipline knew that the positive results Bem reported were somehow simply *wrong*.

Unfortunately for Bem, for while scientific journals (especially those in the social science) have been historically reluctant to publish replications unless the original research was sufficiently controversial, interesting, or counterintuitive, this, too, was beginning to change. And his series of experiments definitely qualified on two (and for some all three) accounts anyway.

Two teams of researchers (Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012) both quickly performed multiple replications of Bem's most impressive studies (the Galak team choosing Bem's eighth and ninth and the Ritchie Wiseman, and French teams his eighth study) and promptly submitted their papers for publication. The Galak group submitted to the same journal that had published Bem's original series (i.e., *The Journal of Personality and Social Psychology*), and, incredibly to some (but not so much to others), the editor of that journal promptly rejected the paper on the basis that it was his journal's policy not to publish

replications. According to Ed Yong (2012), the Ritchie team encountered the same resistance in *Science* and *Psychological Science,* which both said that they did not publish "straight replications." A submission to the *British Journal of Psychology* did result in the paper being sent out for peer review but it was rejected (although Bem having been selected as one of the peer reviewers surely couldn't have influenced that decision) before *PLoS ONE* finally published the paper.

Now, as previously mentioned, it is not known whether this episode marked the birth of the reproducibility movement or was simply one of its several inaugural episodes. And whether it will have an effect on the course of scientific progress (or simply serve as another publishing opportunity for academics), my Bronx mentor will not permit me to guess. But surely the following article is one of the most impactful replications in this very unusual and promising movement (with kudos also to the less cited, but equally impressive, Ritchie et al. replication).

Correcting the Past: Failures to Replicate Psi

Jeff Galak, Robyn A. LeBoeuf, Leif D. Nelson, and Joseph P. Simmons (2012)

Since it took Bem a decade to conduct his nine studies, it is perhaps not surprising that these authors chose to replicate only two of them. Both dealt with the "retroactive facilitation of recall," and the replicating authors reported choosing them because they were (a) the "most impressive" (both were statistically significant and the ninth experiment [as is customary] reported the largest effect size), and (b) the findings in the other seven studies employed affective responses that were more difficult to reproduce and might have been, by then, time-specific (no pun intended).

In all, seven replications were conducted, four for Bem's eighth experiment and three for the ninth. Undergraduates were used in three studies and online participants in the other four. Additional methodological advantages of the replications included:

1. They all used predetermined sample sizes and all employed considerably more participants than Bem's two studies. As previously

discussed, the crucial importance of (a) employing sufficiently large sample sizes for ensuring adequate statistical power and (b) the a priori decision of deciding how many participants to employ avoid two of the most virulent QRPs.

2. The replications used both identical and different words (categories of words) to Bem's. This was apparently done to ensure both that (a) the studies were direct replications of the originals and (b) there wasn't something unusual about Bem's choice of words (a replicator's version of both "having one's cake and eating it, too").
3. There was less contact between research staff and participants in the replications, which is important because the former can unconsciously (or in some cases quite consciously) cue responses from the latter.
4. Post-experimental debriefing included the following question for online samples: "Did you, at any point during this study, do something else (e.g., check e-mail)?" Participants were assured that their answer would not influence their payments for participating. This was done to help ensure that respondents were taking their task seriously and following the protocol. If they were not, then obviously the original findings wouldn't replicate.
5. At least two of the four relevant QRPs that were responsible for producing false-positive results modeled by Simmons, Nelson, and Simonsohn (2011) and *may* have characterized Bem's experiments were completely avoided in the seven replications. One involved choosing when to stop running participants and when to add more (the replicators did not look at their data until the end of the experiments, whereas Bem apparently did and adjusted his procedures accordingly as he went along based on how things were going). (This obvious QRP was reported by a former research assistant in an excellent article on the subject in *Slate Magazine* [Engber, 2017].) The other avoided QRP involved not choosing which dependent variables to report on. (Bem reputably conducted multiple analyses but emphasized only the statistically significant ones).

Of course Bem's original results did not replicate, but our protagonists (Galak, LeBoeuf, Nelson, and Simmons) weren't quite through. For while they had apparently (a) repeated Bem's analyses as closely as possible,

(b) used more participants than he, and (c) ensured that certain QRPs did *not* occur in their replications, who is to say that they themselves might be wrong and Bem himself might be correct?

So, the replicating authors went one step farther. Through an exhaustive search of the published literature they located 10 independent replications of Bem's most impressive two experiments (i.e., numbers 8 and 9) other than their own (of which, it will be recalled, there were seven). Numerically, five of these were in a positive *direction* (i.e., produced a differential between recalled words that were reinforced after the recall test vs. words that were not reinforced), and five favored the words not reinforced versus those there were. (In other words, the expected result from a coin flipping exercise.)

Next our heroes combined all 19 studies (i.e., the 2 by Bem, the 7 by the authors themselves, and the 10 conducted by other researchers) via a standard meta-analytic technique. Somewhat surprisingly, the results showed that, as a gestalt, there was no significant evidence favoring reverse causation even while including Bem's glowingly positive results. When the analysis was repeated using only *replications* of Bem's work (i.e., without including his work), the evidence was even more compelling against psi—reminiscent of Langmuir's description of those "things that aren't so."

So, was the issue settled? Temporarily perhaps, but this and other nonsense resurfaces every few years and surveys of young people who are deluged with (and enjoy watching) screen adaptations of Marvel and DC comic books, devotees of conspiracy theories on YouTube, and adults who continue to frequent alternative medical therapists in order to access the placebo effect will probably continue to believe in the paranormal until the planet suffers a major meteor strike.

And, as a footnote, obviously Bem—like Rhine before him—remained convinced that his original findings were correct. He even

1. Conducted his own meta-analysis (Bem, Tressoldi, Rabeyron, & Duggan, 2016) which, of course, unlike Galak et al.'s was positive;
2. Preregistered the protocol for a self-replication in a *parapsychology registry* (which I had no idea existed) but which theoretically

prevented some of the (probable) original QRPs (http://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_registry_1016.pdf), which it will be recalled involved (a) deep-sixing negative findings, (b) cherry-picking the most propitious outcome variables, (c) tweaking the experimental conditions as he went along, as Bem (in addition to his former research assistant) admitted doing in the original studies, and (d) abandoning "false starts" that interim analyses indicated were trending in the "wrong" direction (which Bem also admitted doing although he could not recall the number of times this occurred); and, finally,

3. Actually conducted said replication with Marilyn Schlitz and Arnaud Delorme which, according to his preregistration, failed to replicate his original finding. However, when presenting his results at a parapsychological conference, Engber reports a typically happy, pathological scientific ending for our intrepid investigator.

> They presented their results last summer, at the most recent [2016] annual meeting of the Parapsychological Association. According to their pre-registered analysis, there was no evidence at all for ESP, nor was there any correlation between the attitudes of the experimenters—whether they were believers or skeptics when it came to psi—and the outcomes of the study. In summary, their large-scale, multisite, pre-registered replication ended in a failure. [That's the bad news, but there's always good news in pathological science.] In their conference abstract, though, Bem and his co-authors found a way to wring some droplets of confirmation from the data. After adding in a set of new statistical tests, *ex post facto*, they concluded that the evidence for ESP was indeed "highly significant."

Of course, pathological science isn't limited to psychology. In fact, Professor Langmuir's famous colloquium (given his professional interests) was more heavily weighted toward the physical than the social sciences. So let's go back a few decades and revisit a much more famous pathological example generated by QRPs involving the physical science, the desire for fame and fortune, and the willingness to see and engage in "the science of things that aren't so" (which, incidentally, would have constituted a better label than "irreproducible science").

The Possibly Even Odder Case of the Discovery of Cold Fusion

While the cold fusion debacle of some three decades ago may have begun to fade from memory (or was never afforded any space therein by some), it still holds some useful lessons for us today. Ultimately, following a few initial bumps in the road, the scientific response to the event should go down as a success story for the replication process in the physical sciences just as the psi replications did for psychology. But first a bit of background.

Nuclear fusion occurs when the nuclei of two atoms are forced into close enough proximity to one another to form a completely different nucleus. The most dramatic example of the phenomenon occurs in stars (and our sun, of course) in the presence of astronomically (excuse the pun) high temperatures and pressures.

Ironically, this celestial process turns out to be the only mechanism by which the alchemists' dreams of changing lighter (or less rare) elements into gold can be realized since their crude laboratory apparatuses couldn't possibly supply the necessary energy to duplicate the fusion process that occurs at the center of stars or the explosion of a nuclear fusion (hydrogen) bomb. But as Irving Langmuir (pathological science), Robert Park (voodoo science), and even Barker Bausell (snake oil science) have illustrated, all of the laws of science can be subverted by a sufficient amount of ambition, ignorance, and disingenuousness.

So it came to pass on March 23, 1989, that the University of Utah held a press conference in which it was breathlessly announced that two chemists, Stanley Pom and Martin Fleischmann, had invented a method that gave promise to fulfilling the dream of eventually producing unlimited, nonpolluting, and cheap energy using a simple tabletop device that would have charmed any alchemist of centuries past. The apparatus (Figure 5.1) itself generating this earthshaking discovery was comprised of

1. An unimposing, tabletop-sized, insulated container designed to maintain the temperature of the contents therein independently of the outside temperature for relatively long periods of time;
2. Heavy water (an isotope of hydrogen containing a proton and a neutron as opposed to only a single proton), plus an electrolyte to facilitate the flow of electricity through it; and
3. A cathode made of palladium (a metallic element similar to platinum).

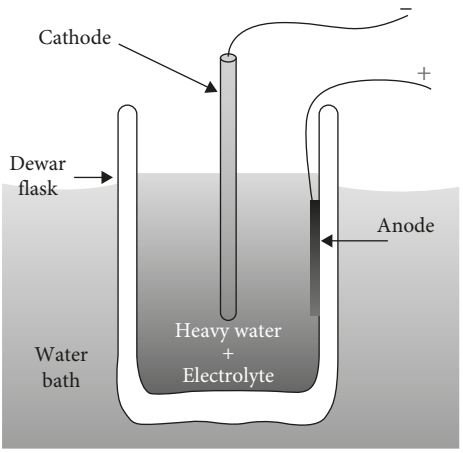

Figure 5.1 The remarkably simple and cheap device that ostensibly produced a cold fusion reaction.
https://en.wikipedia.org/wiki/Cold_fusion#/media/File:Cold_fusion_electrolysis.svg

The press conference occurred before any of these results had been submitted to a peer reviewed journal, apparently to preempt another cold fusion researcher (Stephen Jones) at another Utah university (Brigham Young). Jones's version of inducing nuclear fusion actually had a recognized scientific mechanism of action but was considered completely impractical due to the inconvenient fact (which he apparently recognized) that the tiny amount of energy apparently emanating from his procedures was far exceeded by the amount of energy required to produce it.

In any event, a double-helix style race commenced (although perhaps even more vituperative between Pom and Fleischmann vs. Jones), with Pom suspecting Jones of trying to steal his work (although there is no evidence of this). Unlike the race to characterize the structure of DNA, however, the financial stakes here were so high (potentially involving trillions of dollars) that the University of Utah's press conference was characterized by exaggerated claims about their researchers' actual progress.

As a result, both the press and a large swath of the scientific community appeared to lose their respective minds, with Pom and Fleischmann immediately receiving rock star status accompanied by dozens of laboratories all over the world beginning the process of attempting to *replicate* their results. Very shortly, aided by the simplicity of the intervention, "confirmations" of the experiment were issued by researchers at Georgia

Tech and Texas A&M, but, before long, laboratories at MIT, Cal Tech, Harwell, and others reported failures to do so. To greatly oversimplify the entire drama, along with the back-and-forth accusations and counter claims, Georgia Tech and A&M retracted their confirmations upon re-examining their results.

However, exact replications can be a bit difficult without knowledge of the actual methods employed in the original work, and Stanley Pom (who basically became the spokesperson for the entire fiasco) wasn't about to share *anything* with *anybody* including high-impact peer reviewed journals or scientific competitors. So although there were dozens of attempted replications accompanied by an ever decreasing number of positive results, Pom assured the pathologically gullible press (*The Wall Street Journal* being the primary advocate of cold fusion research, given the trillion-dollar industry it would potentially spawn) that the many failures to replicate could be explained by the simple fact that the labs producing them had not used the exact and proper Pom-Fleischmann procedures (which, of course, was probably true since Pom refused to share those details with them).

But soon, more and more conventional physicist and chemists became increasingly critical (and indeed incredulous) that neither the original Utah experiments nor their "successful" replications bothered to run controls involving, say, plain water instead of its heavy counterpart. For experimental controls are not only *absolute* prerequisites for findings to be considered credible in *any* discipline (social, biological, or physical), but also their absence constitutes an egregious QRP in and of itself. And almost equally important, for some reason, Pom and Fleischmann failed to secure sufficiently sensitive (and available) instrumentation to filter out background environmental contaminants or other sources of laboratory noise. A few sticklers were even concerned that the etiology of the effect violated the first law of thermodynamics (i.e., while energy can be transformed from one form to another, it cannot be created or destroyed, even in a closed system such as our protagonists' tabletop device). But as the old saying goes, "laws are made to be broken."

But counter-arguments such as these were impotent compared to those of a consummate snake oil salesman such as Stanley Pom who dismissed anything with negative implications as unimportant or part of a witch hunt by jealous East Coast physicists, especially those affiliated with MIT and Yale. In fact, Pom was reported by Gary Taubes (1993) in his splendid 473-page history of the incident (entitled *Bad Science: The Short Life and Weird Times*

of Cold Fusion) as saying, when confronted by the ever-growing number of laboratories' failure to find *any* effect, "I'm not interested in negative effects!" (Perhaps he should have been a journal editor.)

And so it went, press conference after press conference, professional conference after professional conference. Positive results were highlighted and negative results were summarily dismissed. Even the advocates who occasionally produced small degrees of excess heat (a supposed indicator of the fusion process but obviously of many other more mundane processes as well) could only do so occasionally. But this was enough to keep the process alive for a while and seemed to add to its mysteriousness. For others, it was obvious that something was amiss, and it wasn't particularly difficult to guess what that was.

What was obvious to Pom, however, was what was really needed: more research, far more funding, partnerships with corporations such as General Electric, and, of course, *faith*. And, every so often, but with increasing rarity, one of the process's proponents would report an extraordinary claim, such as a report from the Texas A&M laboratory that their reaction had produced tritium, a ubiquitous byproduct of fusion. And while this was undoubtedly the most promising finding emanating from the entire episode, like everything else, it only occurred once in a few devices in a single lab which eventually led to a general consensus that the tritium had been spiked by a single individual.

So, gradually, as the failures to replicate continued to pile up and Dr. Pom kept repeating the same polemics, the press moved on to more newsworthy issues such as a man biting a dog somewhere in the heartland. And even the least sensible scientists appeared to have developed a modicum of herd immunity to the extravagant claims and disingenuous pronouncements by the terminally infected. So, in only a very few years, the fad had run its course, although during one of those heady years cold fusion articles became the most frequently published topic area in all of the physical sciences.

Today, research on "hot" (i.e., conventional) fusion continues, the beneficiary of billions in funding, but cold fusion investigations have all but disappeared in the mainstream scientific literature. Most legitimate peer reviewed journals in fact now refuse to even have a cold fusion study reviewed, much less publish one. However, a few unrepentant investigators, like their paranormal and conspiracy theory compatriots, still doggedly pursue the dream and most likely will continue to do so until they die or become too infirm to continue the good fight.

Lessons Learned

Unfortunately, this subheading is an oxymoron when applied to pathological, voodoo, or snake oil science researchers, but it is unlikely that any of these practitioners will ever read a book such as this (if any such exist). The rest of us should always keep Irving Langmuir's lecture on pathological science in mind, especially his six criteria for what makes a scientific discovery pathological, be it cold fusion, N-rays, mitogenetic rays, or photographic evidence proving the existence of flying saucers—the latter of which loosely qualifies as a physical science phenomenon since Langmuir, after examining what investigators considered the "best evidence," concluded that "most of them [extraterrestrial vehicles] were Venus seen in the evening through a murky atmosphere."

So perhaps, with a little effort, present-day scientists can translate at least some of Langmuir's criteria to their own disciplines. Let's use cold fusion as an example.

1. "The maximum effect that is observed is produced by a causative agent of barely detectable intensity, and the magnitude of the effect is substantially independent of the intensity of the cause." This one certainly applies to cold fusion since astronomically large amounts of heat are required to generate nuclear fusion while cold fusion required only a simple, low-voltage electrical current flowing through a liquid medium at room temperature.
2. "The effect is of a magnitude that remains close to the limit of detectability, or many measurements are necessary because of the very low statistical significance of the results." Over and over again responsible physicists unsuccessfully argued that the small amount of heat generated in the tabletop device was millions of times less than that generated by the smallest known fusion reactions. The same was true for the tiny number of emitted neutrons (a byproduct of the process) occasionally claimed to have been measured.
3. "There are claims of great accuracy." As only two examples, the occasional reports of tiny amounts of increased heat as measured by our heroes apparently involved a substandard calorimeter and did not take into account the facts that (a) different solutions result in different degrees of conductivity (hence influencing measurable heat) or even that (b) the amount of commercial electric current (that helped produce

said heat) is not constant and varies according to various conditions (e.g., the performance of air conditioners on hot summer days).

4. "Fantastic theories contrary to experience are suggested." This one is obvious.
5. "Criticisms are met by ad hoc excuses." As well as downright lies, a paranoid belief that the individuals failing to replicate positive findings had hidden or nefarious agendas and/or were not able to employ the original procedures (since, in the case of cold fusion, these happened to be closely guarded secrets). Furthermore, even the most avid supporters of the process admitted that their positive results occurred only sporadically (hence were not reproducible by any scientific definition of the term). Various excuses were advanced to explain this latter inconvenient truth, although only one excuse (admitted ignorance of what was going on) was not disingenuous.
6. "The ratio of supporters to critics rises and then falls gradually to oblivion." As mentioned previously, the number of supporters quickly approached almost epidemic levels, perhaps to a greater extent than for any other pathologically irreproducible finding up to that point. However, facilitated by a discipline-wide replication initiative, the epidemic subsided relatively quickly. But a good guess is that Pom continues to believe in cold fusion as fervently as Bem still believes in psi.

But since Dr. Langmuir was not privy to the cold fusion epidemic or Daryl Bem's landmark discoveries, perhaps he wouldn't object too strenuously ifthree additional principles were added: one advanced by a philosopher of science centuries ago who would have also been a Nobel laureate if the award existed, one attributable to a number of scientific luminaries, and one to completely unknown pundit:

7. "What is done with fewer assumptions is done in vain with more," said William of Occam, who counseled scientists choosing between alternative theories or explanations to prefer the one that required the fewest unproved assumptions.
8. "Extraordinary claims require extraordinary evidence," which, according to Wikipedia was only popularized by Carl Sagan but originally was proposed in one form or another by David Hume, Pierre-Simon Laplace, and perhaps some others as well. A more extraordinary claim is difficult to concoct than the contention that a humble tabletop apparatus could subvert the laws of physics and fuel the world's energy

needs for millennia to come. Or that future events can effect past ones for that matter.

9. When a scientific finding sounds too good (or too unbelievable) to be true, it most likely isn't (pundit unknown).

However, Gary Taubes (1993) probably deserves the final word on the lessons that the cold fusion fiasco has for science in general as well as its direct applicability to the reproducibility crisis.

> Of all the arguments spun forth in defense of cold fusion, the most often heard was *there must be something to it*, otherwise the mainstream scientific community would not have responded so vehemently to the announcement of its discovery. What the champions of cold fusion never seemed to realize, however, or were incapable of acknowledging, was that the vehemence was aimed not at the science of cold fusion, but at the method [i.e., experimental methodology]. Positive results in cold fusion were inevitably characterized by sloppy and amateurish experimental techniques [we could substitute QRPs here]. If these experiments, all hopelessly flawed, were given the credibility for which the proponents of cold fusion argued, the science itself would become an empty and meaningless endeavor. (p. 426)

But Was It Fraud?

As mentioned previously, such judgments are beyond my paygrade. However, one of my three very accomplished virtual mentors (the physicist Robert Park, head of the Washington office of the American Physical Society at the time) was eminently qualified to render such a judgment. So, according to Gary Taubes (to whom I apologize for quoting so often but his book truly is am exemplary exposition of the entire fiasco and should be read in its entirety):

> When the cold fusion announcement was made, Robert Bazell, the science reporter for NBC News, interviewed Robert Park. . . . He asked Park, off the record, whether he thought cold fusion was fraud, and Park said, "No but give it two months and it will be." (p. 314)

And again according to Taubes, Bob was quite proud of this piece of prognostication, even though he admitted that his timeline was off by about 6 weeks.

But how about Daryl Bem and the ever-increasing plethora of positive, often counterintuitive, findings published in today's literature? Is that fraud? Unfortunately, Professor Park is quite ill and not able to grace us with his opinion. So let's just label psi a QRP-generated phenomena and let it go at that. In science, being wrong for the wrong reasons is bad enough.

But Should Outrageous Findings Be Replicated?

"Should" is probably the wrong word here. And anyway, some of us simply don't have the self-control to avoid debunking high-profile, ridiculous, aesthetically offensive (to some of us at least) nonsense. So the higher the profile, the more likely findings are to be replicated.

Of course some topics are politically immune to evidence (think climate change), some are outside the purview of scientific investigation (e.g., religion), and some scientists and non-scientists are so tightly wrapped into their theories or political views that no amount of evidence can change their opinions. Scientists are also humans, after all, so they can be equally susceptible to foibles such as shamelessly hyping the importance of their own work while debunking any conflicting evidence. Whether such behaviors qualify as QRPs is a matter of opinion.

But pathological findings aside, should scientists conduct replications as part of their repertoire? Here, "should" *is* the correct word. First, all scientists should replicate their own work whenever possible, but not as a means of convincing others of their finding's veracity. Scientists in general tend to be skeptical, so many are likely to consider successful self-replications to be too governed by self-interest to be taken at face value. Instead, self-replications should be used as a means of ensuring that one's own work is valid to avoid continuing down a dead end street and wasting precious time and resources. And regardless of whether scientists replicate their findings or not, they should also (a) check and recheck their research findings with an eye toward identifying any QRPs that somehow might have crept into their work during its conduct and (b) cooperate fully with colleagues who wish to independently replicate their work (which, in his defense, Daryl Bem apparently did).

Of course, given the ever increasing glut of new studies being published daily, everything obviously can't be replicated, but when an investigative team plans to conduct a study based on one of these new findings, replication is a sensible strategy. For while a replication is time- and resource-consuming, performing one that is directly relevant to a future project may actually turn out to be a cost- *and* time-saving device. Especially if the modeling results involving the prevalence of false-positive results (e.g., Ioannidis, 2005; Pashler & Harris, 2012) previously discussed have any validity.

However, some studies *must* be replicated if they are paradigmatically relevant enough to potentially challenge the conventional knowledge characterizing an entire field of study. So all scientists in all serious disciplines *should* add the methodologies involved in performing replications to their repertoires.

In the past, "hard" sciences such as physics and chemistry have had a far better record for performing replications of potentially important findings quickly and thoroughly than the "softer" social sciences, but that difference is beginning to fade. As one example, ironically, in the same year (2011) that Daryl Bem published his paradigm-shifting finding, physics experienced a potentially qualitatively similar problem.

In what a social scientist might categorize as a post hoc analysis, the Oscillation Project Emulsion-t Racking Apparatus (OPERA) recorded neutrinos apparently traveling faster than the speed of light (Adam, Agafonova, Aleksandrov, et al., 2011). If true (and not a false-positive observation), this finding would have negated a "cornerstone of modern physics" by questioning a key tenet of Einstein's theory of general relativity.

Needless to say, the physics community was more than a little skeptical of the finding (as, in their defense, were most psychologists regarding the existence of psi), especially given the fact that only seven neutrinos were observed traveling at this breakneck speed. In response, several labs promptly conducted exact replications of the finding within a few months (similar to the speed at which the Galak and Ritchie et al. teams replicated Bem's work). Also not unlike these tangentially analogous social science replications, all failed to produce any hint of faster-than-light travel while concomitantly the original OPERA team discovered the probable cause of the discrepancy in the form of a faulty clock and a loose cable connection. The ensuing embarrassment, partially due to the lab's possibly premature announcement of the original finding, reputably resulted in several leaders of the project to submit their resignations (Grossman, 2012).

So Are the "Hard" Sciences All That Different From Their "Softer" Counterparts?

At present there is little question that the physical sciences are at least more prestigious, successful, and gifted with a somewhat lower degree of publication bias (and thus perhaps more likely to have a lower prevalence of false-positive findings) than the social sciences. The former's cumulative success in generating knowledge, theoretical and useful, is also far ahead of the social sciences although at least some of that success may be due to the former's head start of a few thousand years.

Daniele Fanelli (2010), a leading meta-scientist who has studied many of the differences between the hard and soft sciences, has argued that, to the extent that the two scientific genres perform methodologically comparable research, any differences between them in other respects (e.g., subjectivity) is primarily only "a matter of degree." If by "methodological comparability" Dr. Fanelli means (a) the avoidance of pathological science and (b) the exclusion of the extreme differences in the sensitivity of the measurement instrumentation available to the two genres, then he is undoubtedly correct. However, there seems to be a huge historical affective and behavior gap between their approaches to the replication process which favors the "hard" sciences.

As one example of this differential disciplinary embrace of the replication process, a *Nature* online poll of 1,575 (primarily) physical and life scientists (Baker, 2016) found that 70% of the respondents *had tried and failed* to replicate someone else's study, and, incredibly, almost as many had tried and failed to replicate one of their own personal finding. Among this group, 24% reported having published a successful replication while 13% had published a failure to replicate one, both an interesting twist on the publication bias phenomenon as well as indirect evidence that the replication process may not be as rare as previously believed—especially outside the social sciences.

If surveys such as this are representative of the life sciences, their social counterparts have a long way to go despite the Herculean replication efforts about to be described here. However, it may be that if the social sciences continue to make progress in replicating their findings and begin to value being correct over being published, then perhaps the "hierarchy of science" and terms such as "hard" and "soft" sciences will eventually become archaic.

The Moral Being (Besides Raising the Strawman of Hard Versus Soft Science)

As mentioned previously (e.g., the classic Watson and Crick example), when announcing an honest-to-goodness new discovery, employing circumspect (or at least cautious) language to announce it is a good practice. Even more importantly, an actual new discovery probably shouldn't even be announced unaccompanied by a sufficiently stringent a priori registered alpha level (perhaps 0.005 in the social sciences) and/or before performing a direct replication of said finding.

So with all this in mind, it is now time to finally begin discussing the replication process in a bit more detail since it remains the most definitive strategy available for ferreting out irreproducible scientific findings. In so doing a somewhat heavier emphasis will be placed on the social sciences since the performance of replications appears to have been a neglected component of their repertoires—at least until the second decade of this century.

References

Adam, T., Agafonova, A., Aleksandrov, A., et al. (2011). Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. *airXiv: 1109.4897v1*.

Baker, M. (2016). Is there a reproducibility crisis? *Nature, 533*, 452–454.

Bem D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425.

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, J. (2016). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research, 4,* 1188.

Engber, D. (2017). Daryl Bem proved ESP is real: Which means science is broken. *Slate Magazine*. https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5*, e10068.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103*, 933–948.

Grossman, L. (2012). Leaders of controversial neutrino experiment step down. *New Scientist*. www.newscientist.com/article/dn21656-leaders-of-controversial-neutrino-experiment-step-down/

Hines, T. (1983). *Pseudoscience and the paranormal: A critical examination of the evidence*. Buffalo, NY: Prometheus Books.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Kaptchuk, T. (1999). Intentional ignorance: A history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine, 72*, 389–433.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–526.

Rhine, J. B. (1934). *Extra-sensory perception*. Boston: Bruce Humphries.

Ritchie, S., Wiseman, R., & French, C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's retroactive facilitation of recall effect. *PLoS ONE, 7*, e33423.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Taubes, G. (1993). *Bad science: The short life and weird times of cold fusion*. New York: Random House.

Wagner, M. W., & Monnet, M. (1979). Attitudes of college professors toward extra-sensory perception. *Zetetic Scholar, 5*, 7–17.

Yong, E. (2012). Bad copy. *Nature, 485*, 298–300.

PART II

APPROACHES FOR IDENTIFYING IRREPRODUCIBLE FINDINGS

6

The Replication Process

Previous chapters have hinted at the key role that the replication process plays in enhancing scientific progress and ameliorating the reproducibility crisis. However, there is little about the scientific process that is either easy or perfect—replication included—so two points should probably be reviewed, one affective and one epistemological.

First, *the affective perspective*: those who replicate a study and fail to confirm the original finding should not expect to be embraced with open arms by the original investigators(s). No scientist wishes to be declared incorrect to the rest of the scientific world and most will undoubtedly continue to believe (or at least defend) the validity of their results. So anyone performing a replication should be compulsively careful in its design and conduct by avoiding any potential glitches or repeating the mistakes of the original. And from both a personal and a scientific perspective, all modern replications should

1. *Be preregistered on a publically accessible website prior to data collection*. And as important as this dictum is for original research, it may be even more important for a replication given the amount of blowback likely to be generated by an offended original investigator whose results failed to replicate—or by her or his passionate defenders—as often or not via social media.
2. *Be designed with considerably more statistical power than the original (preferably 0.90)*. For, as mentioned in Chapter 2, if a replication employs the same amount of power as a typical original study (0.50 for psychological experiments), then it will have only a 50% chance of obtaining statistical significance even if a true effect exists and the original finding was correct.
3. *Follow the original design as closely as possible (with the exception of repeating any questionable research practices [QRPs] therein)* since even the slightest deviation constitutes fodder for a counterattack. Exceptions exist here, but any design or procedural changes should be justified (and justifiable) and preferably informed by pilot work.

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0007

4. *Attempt to engage the original investigators in the replication process as much as possible.* There are at least two reasons for this. First, it will help assure the original investigators that the replication is not meant to be an attack on their integrity, hence some of the virulence of their objections to a disconfirming replication can be deflected by requesting any feedback they might have regarding the design of the proposed replication. (This is not only a professional courtesy but good scientific practice since it may result in an offer to share scientific materials and other key information seldom available in a typical journal publication.)
5. *And, perhaps most importantly of all, to take the time to examine the Open Science Framework website* (https://osf.io) *and its abundance of highly recommended instructions, examples, and information available for both the preregistration and replication processes.*

And second, *from an epistemological perspective*, it is important to remember that failures to replicate can occur for a number of reasons, including

1. The original study's methods were flawed and its results were incorrect,
2. The replicating study's approach was flawed and its results were false,
3. Both studies were incorrect, or
4. The methods or participants used in the second study were substantively different from those used in the first study (hence the replication does not match the original in terms of key conditions).

But to make things a bit murkier in social and behavioral research, it is also always possible that the original study finding could have been correct *at the time* but no longer reproducible because (to quote Roland, the Stephen King character) "the world has moved on." This possibility is especially problematic in the scenario-type, often culture-related studies of which psychology, political science, and economics are so fond and in which undergraduates and/or Amazon's Mechanical Turk participants are almost universally employed.

Said another way, it may be that constantly evolving cultural changes can influence responses to interventions. Or, given the number of brief interventional studies being conducted, participants may become more experienced, sophisticated, and therefore more difficult to blind to group membership. Or become more susceptible to demand characteristics purposefully or

accidentally presented in the experimental instructions provided to them. Or, if the original (or a very similar) study has achieved a degree of notoriety by finding its way into the press or disciplinary textbooks, participants may have been exposed to it, recognize the replication's intent, and consequently respond accordingly.

But setting such quibbles aside, replications remain the best indicators we have for judging the validity of social scientific results. So let's begin by examining the types or genres of replications available to investigators.

Exact Replications

While preferable, exact replications are not typically possible in most social sciences. They can be feasible in disciplines such as chemistry, physics, or (theoretically) animal research in which genetically identical rodents reared under standardized conditions are employed. (The concept itself was first proposed and named [at least to my knowledge] by the previously mentioned classic modeling article by Greenwald, Gonzalez, Harris, and Guthrie [1996] who basically dismissed the possibility of its implementation in research involving human participants.)

Research involving human (and probably the majority of animal) experiments aside, exact replications are possible (and recommended) for modeling studies and analyses involving existing databases designed to tease out relationships among variables. The latter (aka a*nalytic replications*) can (and should) be performed to ascertain if the same results can be reproduced using the original data and code *if* the latter is available with adequate documentation, which, unfortunately, as will be discussed in Chapter 10, appears to be surprisingly rare.

So let's now concentrate on the replication of experimental findings, which leads us to the most recommended genre by just about everyone interested in reproducibility, although all forms of replication have their special charms.

Direct (aka Close) Replications

This form of replication involves employing the same procedures used in the original study of interest including (a) instructions to participants

and experimenters, (b) types of participants, (c) interventions, (d) measures, and (e) statistical analyses. Possible exceptions involve different subject pools, such as the use of Amazon's Mechanical Turk employees in lieu of undergraduates and definitely should involve (a) increased sample sizes if the original study is underpowered (which in the social sciences is more likely than not), (b) the avoidance of QRPs present in the original (and, of course, avoiding new ones in the replication), and (c) more appropriate data analytic techniques if the original's are patently unacceptable.

As mentioned, direct replications normally require information from the original investigators since journal articles do not communicate anything approaching the required level of detail needed to actually replicate the procedures described therein. Experimental materials should also be solicited from the original investigators, as well as feedback regarding the proposed protocol before it is preregistered. (There is currently a robust movement to persuade all original investigators to preregister their materials, data, code, and detailed experimental protocols, which, if successful, will greatly facilitate the replication process.)

From a strict reproducibility perspective, this genre of replication is preferable to those that follow because the closer the replication matches the original study, the more confidence can be had in the bottom-line inference (i.e., the original results replicated or they did not). A change in a replication's procedures (other than the correction of an obvious methodological or statistical flaw) will almost certainly be cited by the original author(s) as the reason for a failure to replicate if one occurs.

Several original investigators, for example, have blamed a failure to replicate (occasionally vituperatively, sometimes reasoned) on the computerized administration of interventions (often using Amazon Turk participants) as a substitute for laboratory presentations employing college students. Since this change in the *presentation* of stimuli sometimes results in subtle changes to the interventions themselves, it is reasonable to question whether such studies can still be classified as *direct* replications. However, it may be more reasonable to question whether an effect has any scientific importance if it is so tenuous and fragile that it can't be replicated if respondents read their instructions rather than listening to a research assistant recite them.

Many if not most of the 151 replications of psychological studies conducted by the Open Science Collaboration and the three "Many Labs"

efforts discussed in the next chapter were conducted using approaches and participants similar to those used in the original studies, hence any such discrepancies were not likely to have contributed to the disappointing failure-to-replicate rate in these initiatives—especially since the results from the first "Many Labs" study found that undergraduate and Amazon Mechanical Turk employees responded similarly to one another in its replications.

In addition, all 151 replications were highly powered and at least as methodologically sound as the studies they replicated. The replications also employed dozens of sites and investigators, thereby reducing the possibility of systematic biases due to settings or individual researchers. Of course it is always possible that subtle changes in the presentation or timing of an intervention (which are often necessary in the translation from laboratory to computer) might affect an outcome, but again, if a finding is this fragile, how likely is it to be relevant to human behavior in the noisy milieu of everyday life?

A registered replication report (more on that later) published in *Perspective on Psychological Science* performed by the team of Alogna, Attaya, Aucoin, and colleagues (2014) provides an interesting perspective on both of these issues (i.e., fragility and "minor" alternations in study procedures) in a replication of a somewhat counterintuitive study on a concept (or theory) referred to as "verbal overshadowing." The original study (Schooler & Engstler-Schooler, 1990) involved a scenario in which all participants watched a video of a simulated bank robbery. One group then verbally described the robber while the other performed an irrelevant task listing US states and capitals.

Attempting to commit something to memory normally facilitates later recall but in this case the participants who verbally described the appearance of the culprit were significantly less successful in identifying said culprit from a mock lineup than the comparison group who performed an irrelevant task instead.

The replication study initially failed to support the original finding, but its first author (Jonathan Schooler) objected to a timing change between the two events employed in the replication. Accordingly, the replication team repeated that aspect of the study and the original effect reached statistical significance, although, as usual (Ioannidis, 2008), the initial effect size was larger than the replicated one. (Hence, if nothing else this represents a positive case study involving cooperation between replicators and original investigators.)

As for computer-based versus laboratory-based differences between replications and original studies, the jury remains out. For example, in a response to Hagger, Chatzisarantis, Alberts, and colleagues (2016) failure to replicate something called the "ego depletion effect," the original investigators (Baumeister & Vohs, 2016) argued that "the admirable ideal that all meaningful psychological phenomena can be operationalized as typing on computer keyboards should perhaps be up for debate" (p. 575)—an argument that Daryl Bem and others have also used to suggest a reason for their results' failures to replicate. (Recall that a "failure to replicate" here is meant to represent a study that was replicated but failed to reproduce the original study's bottom-line result.)

Of course a truly unrepentant curmudgeon might suggest that the idea than any societally meaningful real-world phenomena that can be discovered in an academic psychological laboratory employing undergraduate psychology students "should perhaps be up for debate" as well. By way of example, returning to the successful cooperative "verbal overshadowing" example following the tweaking of the time interval separating the video from the pictorial lineup, one wonders how likely this finding would translate to a real-life "operationalization" of the construct? Say, to someone (a) actually witnessing a real armed bank robbery in person (possibly accompanied by fear of being shot), followed by (b) questions regarding the appearance of the robbers anywhere from a few minutes to a few hours later by police arriving on the scene, and then (c) followed by a live police lineup several days or weeks later?

Conceptual (aka Differentiated, Systematic) Replications

This genre of replication is more common (at least in the social sciences) than direct replications. Different authors have slightly different definitions and names for this genre, but basically *conceptual replications* usually involve purposefully changing the intervention or the outcome measure employed in the original study in order to extend the concept or theory guiding that study. (The experimental procedures may also be changed as well, but the underlying purpose of this type of study is normally not to validate the original finding since it is tacitly assumed to be correct.)

Of course different investigators have different objectives in mind for conducting a conceptual replication such as

1. Supporting the original study by extending the conditions and circumstances under which its effect occur (similar to the "external validity" concept),
2. Determining whether a concept or construct replicates from one theoretical arena to another using the same or a similar paradigmatic approach, or (less commonly), and
3. Ascertaining if the original effect is an artifact of the specialized manner in which the study was designed or conducted.

Differing objectives such as these (coupled with investigator expectations and attendant [possibly unconscious] design decisions) tend to dilute the primary advantage of their direct counterparts: namely, to determine whether or not original inferential results replicated. In addition, some reproducibility experts (e.g., Pashler & Harris, 2012) consider conceptual replications to actually increase the incidence of false-positive results while others (e.g., Lindsay & Ehrenberg, 1993) argue that direct replications are primarily "important early in a research program to establish quickly and relatively easily and cheaply whether a new result can be repeated at all." After which, assuming a positive replication, conceptual (aka "differentiated") replications are indicated "to extend the range of conductions under which the [original] result . . . still holds" (p. 221).

Nosek and Lakens (2014) list the following three advantages of direct replications along with a quibbling aside or two on my part which hopefully the authors will forgive:

1. "First, direct replications add data to increase precision of the effect size estimate via meta-analysis" (p. 137).
2. "Second, direct replication can establish generalizability of effects. There is no such thing as an exact replication. [Presumably the authors are referring to psychological experiments involving human participants here.] Any replication will differ in innumerable ways from the original. . . . Successful replication bolsters evidence that all of the sample, setting, and procedural differences presumed to be irrelevant are, in fact, irrelevant." It isn't clear why a "successful," methodologically sound conceptual replication wouldn't also "establish generalizability," but I'll defer here.
3. "Third, direct replications that produce negative results facilitate the identification of boundary conditions for real effects. If existing theory

> anticipates the same result should occur and, with a high-powered test, it does not, then something in the presumed irrelevant differences between original and replication could be the basis for identifying constraints on the effect" (p. 137). William of Occam might have countered by asking "Wouldn't an at least equally parsimonious conclusion be that the theory was *wrong*?"

However, another case *against* conceptual replications (or perhaps even calling them "replications" in the first place) is made quite succinctly by Chris Chambers and Brian Nosek in the previously cited and very informative article by Ed Yong (2012):

> From Chambers: "You can't replicate a concept. . . . It's so subjective. It's anybody's guess as to how similar something needs to be to count as a conceptual replication." [He goes on to illustrate via a priming example how the practice also produces a "logical double-standard" via its ability to verify but not falsify . . . thereby allowing weak results to support one another."] And Brian Nosek adds in the same article that conceptual replications are the "scientific embodiment of confirmation bias. . . . Psychology would suffer if it [the conceptual replication process] wasn't practiced but it doesn't replace direct replication. *To show that "A" is true, you don't do "B." You do "A" again* [emphasis added because this is one of the most iconic quotes in the replication literature]." (p. 300)

While I was initially unmoved by these opinions, after some thought buttressed by the iconic simulation by Simmons, Nelson, and Simonsohn (2011), I have come to believe that the etiology of many conceptual replications involves investigators being seduced by the potential of an exciting, highly cited article and then fiddling with the applicable QRPs to ascertain if a desirable p-value can be obtained with a sufficient plot twist to avoid the study being considered a direct replication. If "successful," the resulting underpowered conceptual replication is published; if not it is deep-sixed. (To go one step farther, it may be that many researchers consider the ability to produce a statistically significant study to be the primary indicator of scientific skill rather than discovering something that can stand the test of time or actually be useful—but this is definitely an unsupported supposition.)

Replication Extensions

Since direct replications have historically been difficult to publish and are perceived by some investigators as less interesting than "original research," one strategy to overcome this prejudice may be to conduct a hybrid extension or two accompanying a direct replication. This is especially feasible in disciplines such as psychology in which multiple experiments are typically included in the same publication, with the first study often involving a direct self-replication of a previously published study. In that case, the remaining experiments could involve conceptual replications thereof in which certain facets of the original intervention, experimental procedures, or outcome variables are changed.

However, Ulrich Schimmack (2012) makes a compelling case against psychological science's love affair with multiple studies in the same publication (think of Daryl Bem's nine-study article in which all but one supported his underlying hypothesis). He demonstrates, among other things, that a surplus of low-powered, statistically significant studies in the same publication can be shown to be

1. Highly improbable (via his "Incredibility-Index"),
2. Less efficient than employing these combined sample sizes into the effort's focal study, and
3. Most likely non-replicable due to their probable reliance on QRPs.

As an example, using a typical power level of 0.50 in psychology (recall that 0.80 is the minimum recommended level), the probability of all five experiments in a five-experiment article reporting statistical significance at the 0.05 level would itself be less than 0.05 even if all five results were in fact "true" positive effects. This is comparable to the probability of obtaining five "heads" on five consecutive coin flips when "heads" was prespecified. Alternatively, the probability of a perfect set of 10 statistically significant studies occurring with a power of 0.50 would be extremely improbable ($p < .001$). And, of course, if some one or more of these experiments reported a p-value substantively less than 0.05, then these two probability levels (i.e., $< .05$ or $< .001$) would be even lower. So perhaps an astute reader (or peer reviewer) should assume that something untoward might be operating here in these multiexperiment scenarios? Perhaps suspecting the presence of a QRP or two? For a more thorough and slightly more technical explication of these

issues, see Ulrich Schimmack's 2012 article aptly entitled "The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles" or the Uri Simonsohn, Leif Nelson, and Joseph Simmons (2014) prescient article, "P-Curve: A Key to the File-Drawer."

Partial Replications

Of course some studies are impossible, too difficult, or too expensive to replicate in their entirety. An example resides in what is probably the best known series of psychological experiments yet conducted (i.e., Stanley Milgram's 1963 obedience studies). In the most famous of these experiments, a confederate encouraged participants to administer electric shocks in increasing doses to another confederate in a separate room who pretended to fail a learning task and was accordingly shocked into unconsciousness.

More than four decades later, and using the original equipment, Jerry Burger (2009) replicated this fifth and most famous experiment in this series up to the point at which the second participant's first verbal complaint occurred. This early stoppage in the replication was necessitated by the likelihood that no modern institutional review board (IRB) would approve the original protocol, in which the participants were subjected to an untoward amount of stress based on the "pain" they were administering to the ostensibly slow "learners." (Burger in turn justified this termination point based on the fact that 65% of the participants in the original experiment continued to administer electric shocks all the way to the end of the fake generator's range, at which point the second confederate was ostensibly unconscious.)

His conclusions, following appropriate caveats, were that

> Although changes in societal attitudes can affect behavior, my findings indicate that the same situational factors that affected obedience in Milgram's participants still operate today. (p. 9)

However it is hard to imagine that many educated people by the turn of this century had not heard of Milgram's obedience study, thereby creating a powerful demand effect. (The only relevant screening procedure employed to ascertain this problem was the exclusion of participants who had taken *more than two* psychology courses. Surely those with one or two psychology courses would have heard of these experiments.)

However, regardless of opinions surrounding the validity of either the original study or its replication, Burger's study is a good example of how a little creativity can provide at least a partial replication when a direct one is infeasible—whether for ethical or practical reasons.

Hypothetical Examples of Direct, Conceptual, Extension, Independent, and Self-Replications

Returning to our hypothetical, long-ago graduate student, let's pretend that he and his advisor were wont to replicate their more exciting (at least to them) findings to ensure their reproducibility (although that word wasn't used in that context in those days). Their motivations would have been lost to time and memory by now if this wasn't a hypothetical scenario, but since it is, let's ascribe their motives to a combination of (a) insecurity that the original findings might not have been "real" (after all, our protagonist was a graduate student) and (b) an untoward degree of skepticism, an occasionally useful scientific trait (but always a socially irritating and not a particularly endearing scientific one).

Let's pretend that our duo's first foray into the replication process involved a relatively rare conceptual replication of a pair of negative studies conducted by a famous educational researcher at the time (James Popham, 1971) who was subsequently served as president of the American Educational Research Association. Jim's original purpose was to find a way to measure teaching proficiency on the basis of student achievement but his studies also demonstrated that trained, experienced teachers were no more effective than mechanics, electricians, and housewives in eliciting student learning. (Some may recall the hubbub surrounding the failure of value-added teacher evaluations [Bausell, 2010, 2017] based on student test scores of a decade or so ago which later proved to be a chimera.)

Naturally any educational graduate student (then or now) would have considered such a finding to be completely counterintuitive (if not demented) since everyone knew (and knows) that schools of education train teachers to produce superior student learning. However, since all doctoral students must find a dissertation topic, let's pretend that this hypothetical one convinced his advisor that they should demonstrate some of the *perceived* weaknesses of Popham's experimental procedures. Accordingly, a conceptual replication (also a term not yet introduced into the professional lexicon) was

performed comparing trained experienced classroom teachers with entry-level undergraduate elementary education students. (The latter participants were selected based on their self-reported lack of teaching experience and were known to have no teacher training.)

Alas, to what would have been to the chagrin of any educational researchers, the replication produced the same inference as Popham's original experiments (i.e., no statistically significant learning differences being obtained between the experienced and trained classroom teachers and their inexperienced, untrained undergraduate counterparts). However, being young, foolish, and exceedingly stubborn, our hero might have decided to replicate his and his advisor's own negative finding using a different operationalization of teacher experience and training involving a controlled comparison of tutoring versus classroom instruction—*an extremely rare self-replication of a replication of a negative finding*.

So, in order to replicate their conceptual replication of their negative teacher experience and training study, they might have simply added that comparison as a third factor in a three-way design producing a 2 (tutoring vs. classroom instruction) by 2 (undergraduate elementary education majors who had completed no mathematics instructional courses or teaching experience vs. those who had received both) by 3 (high, medium, and low levels of student ability based on previous standardized math scores).

Naturally, like all hypothetical studies this one was completed without a hitch (well, to be more realistic, let's pretend that it did possess a minor glitch involving a failure to counterbalance the order of instruction in one of the several schools involved). Let's also pretend that it produced the following results for the replication factor (of course, everyone knows [and had known for a couple of millennia] that tutoring was more effective than classroom instruction):

1. No difference (or even a trend) surfaced between the learning produced by the trained and experienced student teachers and their beginning level undergraduate counterparts, and
2. No interaction (i.e., differential learning results produced) occurred between the two types of teachers within the tutoring versus classroom interventions.

Now, even the most skeptical of graduate students (real or imagined) would probably have been convinced at this point that their original negative

findings regarding teacher training were probably valid—at least within the particular experimental contexts employed. Accordingly, this particular graduate student and his advisor might have abandoned the lack of teacher training/experience as a viable variable in their program of research and instead conducted a single-factor study designed to both replicate the originally positive effect for tutoring and extend it to differently sized instructional groups. Diagrammatically, this latter design might have been depicted as in Table 6.1.

The first two cells (tutoring vs. classroom instruction) involved a near perfect *direct* replication of the factorial study's tutoring versus classroom instruction while the final two cells constituted an *extension* (or, in today's language, a *conceptual* replication) thereof. (The construct underlying both replications was conceptualized as *class size*.) The "near perfect" disclaimer was due to the hypothetical addition to the study outcome of two instructional objectives accompanied by two items each based on said objectives in the hope that they would increase the sensitivity of the outcome measure.

For the results of the single-factor study to have been perfect (a) the direct replication of tutoring versus classroom instruction would have reproduced the investigators' original finding (i.e., the tutored students would have learned more than their classroom counterparts), and (b) the conceptual replications would have proved statistically significant in an incrementally ascending direction as well. To make this myth a bit more realistic (but still heartening), let's pretend that

1. Tutoring was significantly superior to classroom instruction (defined as one teacher to 23 students or 1:23) as well as to the 1:2 and 1:5 small group sizes,
2. Both the 1:2 and 1:5 groups learned significantly more than students taught in a 1:23 classroom setting, but
3. There was no statistically significant difference (other than an aesthetically pleasing numerical trend in the desired direction) between 1:2 and 1:5.

Table 6.1 Direct replication combined with an extension (aka conceptual) replication

Classroom Instruction	Tutoring	2-Student Small Group Instruction	5-Student Small Group Instruction

But would these latter findings replicate? And herein resides a nuanced issue and one of the reasons this allegory was presented. The finding probably would replicate if the same instructional objectives, students with similar instructional backgrounds (i.e., who had not been exposed to the experimental curriculum), and the same amounts of instructional time were all employed.

However, if a completely different instructional unit were substituted (e.g., some facet of reading instruction or even a different mathematical topic), the effect might not have replicated since the original unit was specifically chosen (and honed based on pilot studies) to be capable of registering learning within a classroom setting within the brief instructional time employed. It was not honed to produce superiority between tutoring, small group, and classroom instruction, which arguably would have constituted a QRP (a term which also hadn't been yet coined and which the hypothetical graduate student would have probably simply called "stupid"), but rather to ensure that the subject matter and the test could detect learning gains within the 1:23 comparison group sans any significant "ceiling" or "basement" effects.

Alternately, if the same experimental conditions had been employed in a replication involving 60 minutes of instruction, the effect might not have replicated because too many student in all four conditions might have performed close to the maximum score possible (and, of course, substantively less instructional time would have produced the opposite reason for a failure to replicate). Or if the instructional period had been extended to several weeks and regular classroom teachers had been employed in lieu of supervised undergraduates, the original effect might not have replicated given the reality of teacher noncompliance, which has long bedeviled educational research studies.

So the purpose of these hypothetical examples was to simply illustrate some of the complexities in performing and interpreting different types of replications.

A note on independent versus self-replication: As previously mentioned there is no question that replications performed by independent investigators are far more credible than those performed by the original investigators—if nothing else because of the possibility of fraud, self-interest, self-delusion, and/or the high prevalence of unreported or unrecognized QRPs. Makel, Plucker, and Hegarty (2012), for example, in a survey of more than a century of psychological research found that replications by the same team resulted in a 27% higher rate of confirmatory

results than replications performed by an independent team. And incredibly, "when at least one author was on both the original and replicating articles, only three (out of 167) replications [< 2%] failed to replicate *any* [emphasis added] of the initial findings" (p. 539). Also, using a considerably smaller sample (67 replications) and a different discipline (second-language research), Marsden, Morgan-Short, Thompson, and Abugaber (2018) reported a similar finding for self-replications (only 10% failed to provide any confirmation of the original study results).

So the point regarding self-replications is? Self-replications are most useful for allowing researchers to test the validity of their *own* work in order to avoid (a) wasting their time in pursuing a unprofitable line or inquiry and (b) the embarrassment of being the subject of a negative replication conducted by someone else. But, unfortunately, the practice is too fraught with past abuses (not necessarily fraudulent practices but possibly unrecognized QRPs on the part of the original investigators) to provide much confidence among peer reviewers or skeptical readers. And perhaps self-replications (or any replications for that matter) should not even be submitted for publication in the absence of a preregistered protocol accompanied by an adequate sample size.

How Can Investigators Be Convinced to Replicate the Work of Others?

After all, myriad calls for more replications have historically been to no avail (e.g., Greenwald, 1975; Rosenthal, 1991; Schmidt, 2009). In fact the publication rates in some literatures comprise 2% or less (e.g., Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007, in marketing research; Makel, Plucker, & Hegarty, 2012, in psychology; Makel & Plucker, 2014, in education)—the latter study being the standard-bearer for non-replication at 0.13% for the top 100 educational journals.

True, there have been a number of replications overturning highly cited (even classical) studies, examples of which have been described decades ago by Greenwald (1975) and more recently in Richard Harris's excellent book apocalyptically titled *Rigor Mortis: How Science Creates Worthless Cures, Crushes Hope, and Wastes Billions* (2017). But despite this attention, replications have remained more difficult to publish than original work—hence less attractive to conduct.

However, to paraphrase our Nobel Laureate one final time, things do appear to be changing, as witnessed by a recently unprecedented amount of activity designed to actually conduct replications rather than bemoan the lack thereof. The impetus for this movement appears to be a multidisciplinary scientific anxiety regarding the validity of published scientific results—one aspect of which involves the modeling efforts discussed previously. This fear (or belief) that much of the scientific literature is false has surfaced before (e.g., in the 1970s in psychology), but this time a number of forward-looking, methodologically competent, and very energetic individuals have made the decision to do something about the situation as chronicled by this and previous chapters' sampling of some very impressive individual initiatives involving a sampling of high-profile, negative replications of highly questionable constructs such as psi and priming, poorly conceived genetic studies, and high-tech, extremely expensive fMRI studies.

All of which have "gifted" our scientific literatures with thousands (perhaps tens of thousands) of false-positive results. But a propos the question of how scientists can be convinced to conduct more replications given the difficulties of publishing their findings, let's move on to a discussion of some of the solutions.

Registered Replication Reports

The registered report process (as mentioned in Chapter 1) was primarily designed to decrease publication bias by moving the peer review process from the end of a study to its beginning, thus procedurally avoiding the difficulties of publishing negative results. This basic concept has been extended to the replication process in the form of a *registered replication report* (RRR) in which publication–nonpublication decisions were made based almost exclusively on the replication protocol.

A number of journals have now adopted some form of this innovation; one of the first being *Perspectives on Psychological Science*. So let's briefly examine that journal's enlightened version as described by Simons, Holcombe, and Spellman (2014). Basically, the process involves the following steps:

1. A query is submitted to the journal making "a case for the 'replication value' of the original finding. Has the effect been highly influential? Is it methodologically sound? Is the size of the effect uncertain due

to controversy in the published literature or a lack of published direct replications?" (p. 552).

2. If accepted, "the proposing researchers complete a form detailing the methodological and analysis details of the original study, suggesting how those details will be implemented in the replication, and identifying any discrepancies or missing information" (p. 553). Therein follows a back-and-forth discussion between the replicating proposers, the original author(s) of the research to be replicated, and the editors, which, if promising, results in a formal proposal. How this actually plays out is unclear but most likely some objections are raised by the original investigator(s) who hopefully won't have the last word in how the replication study will be designed and conducted.
3. Since the RRR normally requires the participation of multiple laboratories due to the necessity of recruiting large numbers of participants quickly (as well as reducing the threat of replicator bias), the journal facilitates the process by putting out a call for interested participants. This step is not necessarily adopted by all journals or replicators who may prefer either to select their own collaborators or perform the replication themselves using a single site.
4. But, back to the *Perspectives on Psychological Science* approach, the selected laboratories "document their implementation plan for the study on OpenScienceFramework.org. The editor then verifies that their plan meets all of the specified requirements, and the lab then creates a registered version of their plan. The individual labs must conduct the study by following the preregistered plan; their results are included in the RRR regardless of the outcome" (p. 553).
5. All sites (if multiple ones are employed) employ identical methods, and results accruing therefrom are analyzed via a meta-analytic approach which combines the data in order to obtain an overall p-value. Each participating site's data is registered on the Open Science Framework repository and freely available for other researchers to analyze for their own purposes.

While there are significant advantages to a multiple-lab approach, single-site replications should also be considered if they adhere to the designated journal's specified replication criteria. Obviously what makes this RRR initiative appealing to replicating investigators resides in the commitment made by journal editors that, if the replication is carried off as planned, the

resulting report will be published by their journals regardless of whether the original study results are replicated or not. However, as promising as this innovation is, another one exists that is even more impressive. .

What's Next?

Now that the basics of the replication process has been discussed, it is now time to examine the most impressive step yet taken in the drive for increasing the reproducibility of published research. For as impressive as the methodological contributions discussed to this point have been in informing us of the existence, extent, etiology, and amelioration of the crisis facing science, it is now time to consider an even more ambitious undertaking. Given that the replication process is the ultimate arbiter of scientific reproducibility, a group of forward-thinking and energetic scientists have spearheaded the replication of large clusters of original findings. And it is these initiatives (one of which involved replicating 100 different experiments involving tens of thousands of participants) that constitute the primary subject of Chapter 7.

References

Alogna, V. K., Attaya, M. K. Aucoin, P., et al. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*, 556–578.

Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science, 11*, 574–575.

Bausell, R. B. (2010). *Too simple to fail: A case for educational change*. New York: Oxford University Press.

Bausell, R. B. (2017). *The science of the obvious: Education's repetitive search for what's already known*. Lanham, MD: Rowman & Littlefield.

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist, 64*, 1–11.

Evanschitzky, H., C., Baumgarth, Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business, 60*, 411–414.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology, 33*, 175–183.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573.

Harris, R. (2017). *Rigor mortis: How sloppy science creates worthless cures, crushes hope and wastes billions*. New York: Basic Books.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648.

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *American Statistician, 47*, 217–228.

Makel, M. C., Plucker, H. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science, 7*, 537–542.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*, 304–316.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning, 68*, 321–391.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–378.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–526.

Popham, W. J. (1971). Performance tests of teaching proficiency: Rationale, development, and validation. *American Educational Research Journal, 8*, 105–117.

Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–39). Newbury Park, CA: Sage.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143*, 534–547.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports. *Perspectives on Psychological Science*, *9*, 552–555.

Yong, E. (2012). Bad copy. *Nature, 485*, 298–300.

7

Multiple-Study Replication Initiatives

As crucial as individual replications are to the scientific process, future scientists may look back on the first two decades of this century and conclude that one of the greatest impacts made by the reproducibility initiative involved the replication of sets of studies—hundreds of them, in fact, often conducted by teams of researchers conducted in multiple cooperating laboratories.

This chapter reviews some of these initiatives, along with the lessons they provide. Lessons related not only to the conduct of replications, but for safeguarding the integrity of science itself. So let's begin with, perhaps surprisingly, two large-scale corporate replication efforts.

Preclinical Medical Research

Many scientists may have been unaware of the fact that biotech companies have been replicating promising published preclinical results for some time, especially those published by university laboratories. Anecdotally, an "unspoken rule" in the early venture capital industry is reported to be that at least 50% of preclinical studies, even those published in high-impact academic journals, "can't be repeated with the same conclusions by an industrial lab" (Osherovich, 2011, quoting Bruce Booth, a venture capitalist).

The reproducibility of such studies is especially important for the pharmaceutical industry because preclinical work with cells, tissues, and/or animals form the basis for the development of new clinical drugs. And while this preliminary research is costly, it pales in comparison to the costs of drug development and the large controlled efficacy trials required by the US Food and Drug Administration (FDA) prior to clinical use and marketing.

So if these positive preclinical results are wrong, the drugs on which they are based will almost surely not work and the companies that develop and test them will lose a formidable amount of money. And we all know that

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0008

pharmaceutical companies are very much more interested in making money than losing it.

The Amgen Initiative

And thus here enters Glenn Begley into our drama, stage left (Begley & Ellis, 2012). As a senior researcher in Amgen, a major biotech firm, Dr. Begley, like his counterparts in many such companies, constantly monitored the published literature for preclinical results that might have important clinical implications. And whenever an especially promising one was found, Dr. Begley would either have the published results replicated in his own company's hematology and oncology department labs to see if they were valid (which often they were not) or file the study away for future consideration.

What makes Dr. Begley so central to this part of our story is that, prior to leaving Amgen for an academic position, he decided to "clean up his file cabinet" of 53 promising effects to see if any of them turned out to be as promising as their published findings suggested. (The papers themselves "were deliberately selected that described something completely new, such as fresh approaches to targeting cancers or alternative clinical uses for existing therapeutics" [p. 532].)

Of the 53 studies replicated, the results from only 6 (or 11%) were reproducible. This is obviously a finding with very disturbing implications for this genre of research, medical treatment, and science in general as succinctly explained in the authors' words:

> Some non-reproducible preclinical papers [i.e., of the 53 studies replicated] had spawned an entire field, with hundreds of secondary publications [i.e., conceptual replications] that expanded on elements of the original observation, but did not actually seek to confirm or falsify its fundamental basis. More troubling, some of the research has triggered a series of clinical studies—suggesting that many patients had subjected themselves to a trial of a regimen or agent that probably wouldn't work. (p. 532)

Following Begley's exit from the company, Amgen apparently took steps to continue his initiative. In an interview (Kaiser, 2016) with Bruce Alberts,

former editor-in-chief of *Science Magazine*, Amgen announced the creation of a new online journal (*The Preclinical Reproducibility and Robustness Gateway*) designed to publish failed efforts to replicate other investigators' findings in biomedicine and "is seeding the publication with reports on its own futile attempts to replicate three studies in diabetes and neurodegenerative disease in the hope that other companies will follow suit." The purpose of this initiative, according to Sasha Kamb (Amgen's senior vice president for research) is to reduce wasted time and resources following-up on flawed findings as well as "to help improve the self-correcting nature of science to benefit society as a whole, including those of us trying to create new medicines."

The Bayer Initiative

In another of these "mass" replication efforts, the huge pharmaceutical company Bayer Heath Care performed a replication of 67 published projects as part of their "target identification and validation" program (Prinz, Schlange, & Asadullah, 2011). Comparing their results to the published data (with the cooperation of the 23 heads of the laboratories involved in producing the original studies), the following results were obtained:

> This analysis revealed that only in ~20–25% of the projects were the relevant published data completely in line with our in-house findings. In almost two-thirds of the projects, there were inconsistencies between published data and in-house data that either considerably prolonged the duration of the target validation process or, in most cases, resulted in termination of the projects because the evidence that was generated for the therapeutic hypothesis was insufficient to justify further investments into these projects. (p. 713)

While this is a seminal reproducibility initiative in its own right, it would have been preferable if the authors were to have provided a more precise definition and estimate of reproducibility–irreproducibility than approximately *20–25%* of the replicated data not being completely in line with the original results or that in, "*almost* two-thirds" of the cases the two data sources were inconsistent enough "*in most cases*" to result in termination. However, point estimates in "mass" replication studies such as this are not as important as

their bottom line, which is that "a majority of these potentially important studies failed to replicate."

The Cancer Biology Initiative

Perhaps in response to these rather disheartening studies, another ambitious initiative spearheaded by an amazing organization that will be described shortly (the Center for Open Science, headed by a whirling dervish of a psychologist named Brian Nosek) secured funding to replicate 50 high-impact cancer biology studies. Interestingly, Glenn Begley is reported (Harris, 2017) to have resigned from this particular project because he argues that repeating their abysmally poor design will produce meaningless results even if they do replicate.

Unfortunately, this project has run into a number of other setbacks as of this writing (January 2019). First, the costs were more than anticipated (originally budgeted at $25,000 per study, the actual costs rose to more than $60,000 [Kaiser, 2018]). This, among other problems (e.g., the difficulties of reproducing some of the laboratory materials and the unexpected amount of time necessary to troubleshoot or optimize experiments to get meaningful results), has resulted in reducing the originally planned 60 replications to 37 in 2015 and then further down to 29 as of 2017.

The online open-access journal *elife* has been keeping a running tab of the results, which are somewhat muddled because clear replicated versus non-replicated results aren't as definitive as they would be for a single social science finding involving one intervention and one outcome variable (or a physics experiment measuring the speed of neutrinos). Again, as of early 2019, the following 12 results were reported (https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology):

1. Results were replicated: 4
2. Results were not replicated: 2
3. Some results were replicated and some were not: 4
4. Results could not be interpreted: 2

In the spirit of transparency, after reading each study, I could not definitively categorize results 3 and 4. Hence, a 33% failure-to-replicate in some aspect based upon these four categories are reported in Table 7.1, which summarizes the 11 replication initiatives discussed in this chapter.

Table 7.1 Results of 11 major replication initiatives

Study	# Replications	# Replication failures	% Failures
Amgen	53	47	89%
Bayer	67	44[a]	66%
Preclinical Cancer Biology	6	2	33%
Preclinical Materials	238	109[b]	46%
Psychology (Self-Reported)[c]	257	130	51%
Open Science Collaboration	100	64	64%
Many Labs I	13	3	23%
Many Labs II (*In Press*)	28	13	46%
Many Labs III[d]	10	5	50%
Experimental Economics	18	7	39%
Social Science Research	21	8	38%
Total	811	432	53.3%

[a] As mentioned in the text, this is an estimate since the authors reported that in only 20–25% of the cases were the results of the replication identical, and, in *most* of "*almost two-thirds*" of the cases, the results were not sufficiently close to merit further work. Many reports citing this study report a failure rate of 75–80%, but this seems to ignore the "almost two-thirds" estimate, whatever that means.

[b] Results were not reported in terms of number of studies but instead used the number of resources employed across all 238 studies. Therefore this figure (46%) was applied to the number of studies which would add a (probably relatively) small amount of imprecision to this number.

[c] Hartshorne and Schachner (2012), based on a survey of self-reported replications.

[d] This estimate was based on the nine direct and one conceptual replication and not the added effects or interactions.

Another Unique Approach to Preclinical Replication

One of the lessons learned from the cancer biology project is that the replication of preclinical findings is more involved than in social science research due to the complexity of the experimental materials—exacerbated by laboratories not keeping detailed workflows and the effects of passing time on the realities of fading memories and personnel changes.

This is perhaps best illustrated by Vasilevsky, Brush, Paddock, and colleagues (2013) who examined this very problematic source of error via a rather unique approach to determining reproducibility in laboratory research. While this study does not employ actual replications, its authors (as well as Freedman, Cockburn, & Simcoe [2015] before them) argue that without sufficient (or locatable) and specific laboratory materials (e.g., antibodies, cell lines, knockout reagents) a study cannot be definitively replicated

(hence is *effectively irreproducible*). In fact, Freedman and colleagues advance an interesting definition of irreproducibility that is probably applicable to all research genres. Namely, that irreproducibility:

> Encompasses the existence and propagation of one or more errors, flaws, inadequacies, or omissions (collectively referred to as errors) that prevent replication of results. (p. 2)

So, incorporating this definition, and based on 238 life science studies (e.g., biology, immunology, neuroscience), Vasilevsky et al.'s calculation of the number of unique (or specific) experimental materials that could *not* be identified were 56% of antibodies, 57% of cell lines, 75% of constructs such as DNA synthesized for a single RNA strand, 17% of knockout regents, and 23% of the organisms employed.

The Costs of All of These Irreproducible Results

Well, fortuitously, someone has provided us with what appears to be a reasonable estimate thereof via the following unique analysis of the costs of irreproducibility for one meta-discipline.

The Economics of Reproducibility in Preclinical Research

Leonard Freedman, Iain Cockburn, and Timothy Simcoe (2015)

This is the only study documenting the costs of irreproducibility of which I am aware. Extrapolating from 2012 data, the authors estimate that $56.4 billion per year is spent on preclinical research, of which a little over two-thirds is funded by the government. Assuming that 50% of this research is irreproducible (which is certainly not an unreasonable estimate based on Table 7.1), then $28 billion is spent on irreproducible preclinical research. When this is broken down by category, the authors offer the following estimates of the causes of preclinical reproducibility:

1. Biological reagents and reference materials (36.1%)
2. Study design (27.6%)

3. Data analysis and reporting (25.5%)
4. Laboratory protocols (10.8%)

Using one example from the first category (experimental materials, which you may recall constituted the sole target of the Vasilevsky et. al. [2013] study with an overall failure rate of 46%), the authors use a single component of *that* category (cell lines) to illustrate the severe problems inherent in conducting reproducible preclinical cancer studies:

> An illustrative example [i.e., of the problems involving cell lines in general] is the use and misuse of cancer cell lines. The history of cell lines used in biomedical research is riddled with misidentification and cross-contamination events [Lorsch, Collins, & Lippincott-Schwartz, 2014, is cited here, who incidentally report that more than 400 widely used cell lines worldwide have been shown to have been misidentified] which have been estimated to range from 15% to 36% (Hughes, Marshall, Reid, et al., 2007). Yet despite the availability of the short tandem repeat (STR) analysis as an accepted standard to authenticate cell lines, and its relatively low cost (approximately $200 per assay), only one-third of labs typically test their cell lines for identity. (p. 5)

The authors go on to list a number of potential solutions (some of which are generic to research itself and have been [or will be] discussed here, whereas some are primarily targeted at preclinical researchers and won't be presented). However, their concluding statement is worthy of being chiseled on a stone tablet somewhere and (with a disciplinary word change or two) considered by everyone interested in the integrity and reproducibility of science:

> Real solutions, such as addressing errors in study design and using high quality biological reagents and reference materials, will require time, resources, and collaboration between diverse stakeholders that will be a key precursor to change. Millions of patients are waiting for therapies and cures that must first survive preclinical challenges. Although any effort to improve reproducibility levels will require a measured investment in capital and time, the long term benefits to society that are derived from increased scientific fidelity will greatly exceed the upfront costs. (p. 7)

Experimental Psychology: The Seminal "Reproducibility Project"

The psychological version of the Amgen and Bayer initiatives (but comprised of even more replications reported in much more detail) was first announced in 2012 (Brian Nosek, corresponding author, representing the Open Science Collaboration). The results followed 3 years later in a report targeting the scientific community as a whole and published in *Science Magazine* (again with Brian Nosek as the corresponding author along with a Who's Who of psychological investigators with a special interest in scientific reproducibility).

The following abstract represents an attempt to summarize the scope of this seminal scientific effort and describe some of its more salient results. Naturally, the article itself should be read in its entirety (and probably already has been by the majority of the readers of this book, for whom the abstraction will be a review).

Estimating the Reproducibility of Psychological Science

The Open Science Collaboration (2015)

Somehow, 270 investigators were recruited and convinced to participate in the replication of 100 psychological experiments published in three high-impact psychology journals (*Psychological Science, Journal of Personality and Social Psychology,* and *Journal of Experimental Psychology: Learning, Memory, and Cognition*). The logistics of the effort itself were historically unprecedented in the social sciences, and their successful consummation represented (at least to me) an almost incomprehensible feat.

Unfortunately, it wasn't possible for the sampling of articles to be random since it was necessary to match the studies themselves with the expertise and interests of the replicating teams. In addition, some studies were judged too difficult or expensive to replicate (e.g., those employing such difficult-to-recruit participants as autistic children or that required the availability of specialized and expensive tests such as magnetic resonance imaging [MRI]). While these practical constraints may not have provided a precise disciplinary reproducibility rate (which of course neither did the preclinical initiatives just discussed), the project may

have resulted in a reasonable estimate thereof for brief psychological interventions.

But, these caveats aside, the design and methodology of the replications were nevertheless both exemplary and remarkable. Equally impressive, of the 153 eligible studies available, 111 articles were selected for replication, and 100 of these were actually completed by the prespecified deadline. Fidelity of the replications was facilitated by using the actual study materials supplied by the original investigators. These investigators were also consulted with respect to their opinions regarding any divergences from the original design that might interfere with replicability. The average statistical power available for the 100 replications was in excess of 0.90 based on the originally obtained effect sizes. And, of course, all replications were preregistered.

Overall, the replication effect sizes were approximately 50% less than those of the original studies, and their published statistical significance decreased from 97% for the 100 original studies to 36% in the replications. (A statistically significant reduction of less than 10% would have been expected by chance alone.) In addition, the large number of studies replicated provided the opportunity to identify correlates of replication successes and failures, which greatly expands our knowledge regarding the replication process itself. Examples include the following:

1. The larger the original effect size, the more likely the finding was to replicate [recall that the larger the effect size, the lower the obtained p-value when the sample size is held constant a la the simulations discussed previously], so this finding also extends to the generalization that, everything else being equal, *the lower the obtained p-value the more likely a finding is to replicate*.
2. The more scientifically "surprising" the original finding was (in the a priori opinion of the replication investigators, who were themselves quite conversant with the psychological literature), the *less* likely the finding was to replicate.
3. The more difficult the study procedures were to implement, the less likely the finding was to replicate.
4. Studies involving cognitive psychology topics were more likely to replicate than those involving social psychology.

In their discussion, the authors make a number of important points as well, most notably

1. "It is too easy to conclude that successful replication means that the theoretical understanding of the original finding is correct. Direct replication mainly provides evidence for the *reliability of a result. If there are alternative explanations for the original finding, those alternatives could likewise account for the replication* [emphasis added]. Understanding is achieved through multiple, diverse investigations that provide converging support for a theoretical interpretation and rule out alternative explanations" (p. aac4716-6).
2. "It is also too easy to conclude that a failure to replicate a result means that the original evidence was a false positive. Replications can fail if the replication methodology differs from the original in ways that interfere with observing the effect" (p. aac4716-6). [As, of course, does random error and the presence of one or more questionable research practices (QRPs) performed by the replicators themselves.]
3. "How can we maximize the rate of research progress? Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both" (p. aac4716-7).

The authors clearly identify publication bias (investigator and reviewer preferences for positive results) along with insufficient statistical power as primary villains in the failure of 36% of the original studies to replicate. Inextricably coupled with this, but certainly to a lesser extent, is regression to the mean, given the fact that 97% of the studies reviewed were published as positive (recall how close this was to the earlier estimates [96%] of positive psychology results conducted more than four decades ago).

Many Labs 1

This is another extremely impressive cooperative psychology replication project initiated just a bit later than the Open Science Project (Klein, Ratliff, Vianello, et al., 2014). It involved 36 samples totaling 6,344 participants in

the replication of 13 "classic and contemporary" studies. Of the 13 effects, 11 (85%) replicated, although this may partly be a function of the authors' declaration that "some" of the 13 elected effects had already been replicated and were "known to be highly replicable."

This and the other two "Many Labs" projects employed methodologies similar to the Open Science replications, although in addition to their replications they possessed other agenda which were methodological in nature and involved exploring the possibility that the failure to replicate an effect might be due to factors unique to the process itself. Thus, in this first project, differences in replicating sites and/or type of participants (e.g., online respondents vs. students) were explored to ascertain their relationship (if any) to replication/non-replication. Perhaps not surprisingly, the interventions themselves accounted for substantively more of the between-study variation than the sites or samples employed.

Many Labs 2

This study (Klein, Vianello, Hasselman, et al., 2018) was basically an expansion of the Many Labs 1 effort exploring the effects of variations in replicability across samples and settings. This time, 28 published findings were selected for replication of which 15 (54%) were declared replicable, although the obtained effect sizes were considerably smaller than in the original studies, which is commonly the case in replication research.

To determine variations across samples and settings, "each protocol was administered to approximately half of 125 samples and 15,305 total participants from 36 countries and territories." As in the first study, the variability attributable to the different samples and types of participants was relatively small.

Many Labs 3

In keeping with the Many Labs approach to coupling methodological concerns with experimental replications, the purpose of this study (Ebersole. Athertomb, Belangeret, et al., 2016) was to ascertain if the point in the academic semester at which student participants engaged in an experiment was related to reproducibility. More than 3,000 participants were employed

involving 20 sites in addition to online participants. Counting the conceptual replication, 5 of the 10 studies apparently replicated the originals, yielding a 50% success rate. In general, time during the academic semester in which participation occurred did not surface as a substantive moderating variable.

A Summary of the Methodologies Employed

Both the Open Science and the Many Labs replications (the latter borrowing much of its infrastructure from the former) appear to have been designed and reported as definitively, transparently, and fairly as humanly possible. In addition, these initiatives' procedural strategies are relevant for any future replications of experiments and should be followed as closely as possible: a sampling follows:

1. First, the original investigators were contacted in order to (a) obtain study materials if available and necessary, (b) apprise said investigators of the replication protocol, and (c) obtain any feedback these investigators might have regarding the project. (The latter is both a courtesy to the original investigators and in some cases a necessary condition for conducting a direct replication if the original materials are not otherwise accessible.) The replicating team was not required to accept any suggested changes to the planned protocol, but, generally speaking, the Open Science investigators appeared to accept almost all such feedback when proffered since (again) not everything occurring in a study tends to be reported in a journal article.
2. Next, and most importantly, the design of the replication and the planned analysis were preregistered and any necessary divergences from this plan were detailed in the final report. Both of these steps should occur whenever a replication (or any study for that matter) is published. Not to do so constitutes a QRP (at least for replications occurring after 2016—the publication date for this Open Science report—or perhaps more fairly after 2018 as suggested in Chapter 10).
3. The Open Science replications employed considerably larger sample sizes than the original studies in order to ensure statistical power levels that were at least 0.80 (usually ≥ 0.90). Sufficient power is essential for all replications since inadequate power levels greatly reduce the credibility of any research. (Recall that since the typical statistical power for

a psychological experiment is 0.35 for an effect size of 0.50; hence a replication employing exactly the same sample size would have only a 35% chance of replicating the original study even if the original positive result was valid.)

4. It is a rare psychological research publication that reports only one study (76% of those replicated by the Open Science Collaboration reported two or more) and an even rarer one that reports only a single p-value. To overcome this potential problem the Open Science replicators typically chose the final study along with the p-value reported therein which they considered to be associated with the most important result in that study. (If the original author disagreed and requested that a different effect be selected instead, the replicating investigators typically complied with the request.)
5. The replication results were reanalyzed by an Open Science-appointed statistician to ensure accuracy. This is not a bad idea if the analysis of a replication is performed by a non-statistician and should probably be universally copied—as should the other strategies just listed for that matter.

A brief note on the use of multiple sites in the replication of single studies: Medical research has employed multisite clinical trials for decades due to the difficulty of recruiting sufficient numbers of patients with certain rare diagnoses. Psychological studies are not commonly affected by this problem but they do often require more participants than a single site can supply, hence the discipline has invented its own terms for the strategy such as "crowdsourcing science" or "horizontal versus vertical approaches to science."

However, regardless of terminology, multicenter trials unquestionably have a number of advantages over single-site research, as well as unique organizing and coordinating challenges of their own—some of which are specific to the discipline. The ultimate impact of this approach on the future of psychology is, of course, unknown, but if nothing else high-powered studies (both original and replicates) are considerably more likely to be valid than their underpowered counterparts. And while the use of multiple sites introduces additional sources of variance, these can be handled statistically. (As well, this variance, systematic or erroneous, is normally overwhelmed by the increased sample sizes that "crowdsourcing" makes possible.)

Advice for assembling, designing, and conducting all aspects of such studies from a psychological perspective is clearly detailed in an article

entitled "Crowdsourcing Science: Scientific Utopia III" (Uhlmann, Ebersole, & Chartier, 2019) and need not be delineated here. (Utopias I and II will be discussed shortly.)

A Survey Approach to Tracking Replications

Joshua Hartshorne and Adena Schachner (2012) report an interesting approach to tracking the frequency of replication attempts by psychological researchers. Employing a sample of 100 "colleagues of the authors," 49 responded, reporting a total of 257 replications in answer to the following question:

> Approximately how many times have you attempted to replicate a published study? Please count only completed attempts—that is, those with at least as many subjects as the original study. (Hartshorne & Schachner, 2012, Appendix)

Of the 257 replications performed, only 127 (49%) studies fully validated the original results. And if these self-reports are applicable, this suggests that replications are considerably more common in psychology than generally supposed.

Given the sampling procedure employed, no projections to a larger population are possible (e.g., 14 of the respondents were graduate students), but that caveat applies in one degree or another to all of the multiple replication initiatives presented in this chapter. With that said, some variant of this survey approach could constitute a promising method for tracking replications if the identity of the original studies was to be obtained and some information regarding the replication attempt was available (preferably with sharable data).

Experimental Economics

Colin Camerer, Anna Dreber, Eskil Forsell, et al. (2016) replicated 18 studies published in two prominent economic journals between 2011 and 2014. All replications were powered at the 0.90 level or above, and 11 of the 18 studies replicated, yielding a 61% success rate. (As in the Open Science Collaboration

and the Many Labs, study procedures were preregistered.) Unlike the Open Science initiative, which also employed expert predictions regarding replication results using an economic tool called *prediction markets* (Dreber, Pfeiffer, Almenberg, et al., 2015), the expert predictions in the Camerer et al. effort were not particularly accurate.

Social Science Studies in General

Science and *Nature* are among the highest profile, most often cited, and most prestigious journals in science. They are also the most coveted publishing outlets as perhaps illustrated by the average bounty (i.e., in excess of $40,000) paid by China (Quan, Chen, & Shu, 2017) to any home-grown scientists who are fortunate enough to garner an acceptance email therefrom, although that bounty has apparently been terminated recently.

As such, these journals have the pick of many litters on what to publish, and their tendency appears (perhaps more than any other extremely high-impact journals) to favor innovative, potentially popular studies. However, their exclusiveness should also enable them to publish methodologically higher quality research than the average journal, so these disparate characteristics make their studies an interesting choice for a replication initiative.

Accordingly, in 2018, Colin Camerer, Anna Dreber, Felix Holzmeister, and a team of 21 other investigators (several of whom were involved in the previously discussed replication of 18 economics studies) performed replications of 21 experimental social science studies published in *Science* and *Nature* between 2010 and 2015. The selected experiments were required to (a) report a p-value associated with at least one hypothesis and (b) be replicable with easily accessible participants (e.g., students or Amazon Mechanical Turk employees). The replicating team followed the original studies' procedures as closely as possible, secured the cooperation of all but one of the original authors, and ensured adequate statistical power via the following rather interesting two-stage process:

> In stage 1, we had 90% power to detect 75% of the original effect size at the 5% significance level in a two-sided test. If the original result replicated in stage 1 (a two-sided $P < 0.05$ and an effect in the same direction as in the original study), no further data collection was carried out. If the original result did not replicate in stage 1, we carried out a second data collection

> in stage 2 to have 90% power to detect 50% of the original effect size for the first and second data collections pooled. (p. 2)

Of the 21 findings employed, 13 (or 61.9%) replicated the original results employing p-values. (A number of other analyses and replication criteria were also used, including the omnipresent reduction in effect sizes, this time being slightly greater than 50%.) One rather creative analysis involved meta-analyses of the original and replicated studies, which resulted in an important (although generally known) conclusion: namely, that "true-positive findings will overestimate effect sizes on average—even if the study replicates."

A Summary of the 11 Replication Initiatives

First a tabular summarization of the 11 projects just discussed (Table 7.1). (Note that programmatic replication of *existing datasets* are not included here, such as Gertler, Baliani, and Romero's 2018 failure to find "both raw data and usable code that ran" in 84% of 203 published economic studies or the International Initiative for Impact Evaluation's 2018 considerably more successful program for validating their investigators' datasets https://www.3ieimpact.org/evidence-hub/publications/replication-papers/savings-revisited-replication-study-savings. Note also that the Chabris, Herbert, Benjamin, and colleagues' (2012) failure to replicate individual gene–intelligence associations in Chapter 4 were not included because (a) none of the reported associates replicated (hence would skew the Table 7.1 results) and (b) these studies represent an approach that is not used in the discipline following its migration to genome-wide associations.)

Also not included in the preceding calculations are a "mass" (aka crowdsourced) replication initiative (Schweinsberg, Madana, Vianello, et al., 2016) involving the replication of a set of 10 unpublished psychological studies conducted by a single investigator (Eric Uhlmann and colleagues) centered on a single theoretical topic (moral judgment).

In one sense this replication effort is quite interesting because the investigator of the 10 studies reports that two of the key QRPs believed to be responsible for producing irreproducible results were not present in his original studies: (a) repeated analyses during the course of the study and (b) dropping participants for any reason. The authors consider these methodological steps

to be major factors in the production of false-positive results and avoiding them presumably should increase the replicability of the 10 studies.

In another sense, however, the original investigator's choosing of the replicators and the studies to be replicated is somewhat problematic in the sense that it positions the process somewhere between self- and independent replication efforts. The authors of the study, on the other hand, consider this to be a major strength in the sense that it helped (a) duplicate the original contexts of the 10 studies and (b) ensure the experience of the replicating labs in conducting such studies. In any event, the resulting replications produced positive evidence for the reproducibility of 8 of the 10 originally positive studies (1 of the 2 originally negative studies proved to be statistically significant when replicated whereas the other did not).

All in all it is difficult to classify the positive replications in this effort. They appeared to be well-conducted, highly powered, preregistered, and methodologically sound (i.e., by employing a variant of the "Many Labs" approach). So it may be unfair to exclude them from the Table 7.1 results simply because they were selected in part because they were expected to produce positive replications and the replicating laboratories were personally selected by the original investigator. So, for the hopefully silent majority who wish to take issue with this decision, adding these eight out of eight positive replications of positive studies to Table 7.2 produces the overall results shown in Table 7.2.

Close but still an apples versus oranges comparison and weak support for the Ioannidis and Pashler and Harris modeling results.

And, as always, there are probably additional multiple replication initiatives that escaped my search, hence the list presented here is undoubtedly incomplete. In addition, an impressive replication of 17 structural brain–behavior correlations (Boekel, Wagenmakers, Belay, et al., 2015) was not included because it relied on a Bayesian approach which employed different replication/non-replication criteria from the previous 12 efforts. This study's finding is as follows:

Table 7.2 Revision: Results of 12 major replication initiatives

Study	# Replications	# Replication failures	% Failures
Total	811 + 8	432 + 0	52.7%

> For all but one of the 17 findings under scrutiny, confirmatory Bayesian hypothesis tests indicated evidence in favor of the null hypothesis [i.e., were negative] ranging from anecdotal (Bayes factor < 3) to strong (Bayes factor > 10). (p. 115)

Five Concluding Thoughts Regarding These Multiple Replications

First of all, the Amgen and Bayer preclinical initiatives basically provided no procedural details regarding how the replications were conducted or how replication versus non-replication was operationalized (other than they didn't merit follow-up work). However, these two efforts were of significant importance because they were early multiple replication attempts and their topic area was arguably of greater scientific, economic, and societal import than any of the other disciplinary initiatives except perhaps the troubled biology cancer initiative.

Second, while all 11 of these initiatives were impressive and important contributions to the scientific reproducibility knowledge base, none provides a generalizable estimate of the prevalence of false-positive results in their literatures due to their understandably unsystematic and non-random selection criterion. Recognizing the very real dangers here of combining apples and oranges, the 11 initiatives as a whole involved a total of 811 studies, of which 432 failed to replicate. *This yielded a 53.3% failure-to-replicate rate.* Not a particularly heartening finding but surprisingly compatible with both Ioannidis's and Pashler and Harris's Chapter 2 modeling estimates.

Third, from an overall scientific perspective, the importance of these initiatives is that if 25% were to be considered an acceptable level for irreproducible results, only 1 (the first "many labs" project) of the 11 initiatives reached this level. (And recall that an unspecified number of the "Many Labs I" studies were reported to have been selected *because* they had already been successfully replicated.)

Fourth, although I have reported what amounts to hearsay regarding the details of Glenn Begley's resignation from the Open Science cancer biology initiative designed to replicate 50 high-impact cancer biology studies, I do agree with Professor Begley's reported objection. Namely, that if a study's design and conduct are sufficiently deficient, a high-fidelity replication thereof employing the same QRPs (sans perhaps low statistical power) is

uninformative. And while psychological versus preclinical experiments may differ with respect to the types and prevalence of these artifacts, the end result of such failings will be the same in any discipline: a deck carefully and successfully stacked to increase the prevalence of false-positive results in both original research *and* its replication.

And finally, all of these initiatives are basically exploratory demonstration projects conducted for the betterment of science and for the benefit of future scientists. Furthermore, none of the authors of these papers made any pretense that their truly impressive approaches would solve the reproducibility crisis or even that their results were representative of their areas of endeavor. They have simply taken the time and the effort to do what they could to alert the scientific community to a serious problem for the betterment of *their* individual sciences.

And an apology: there are undoubtedly more multiple-study replications under way than have been discussed here. Engineering, for example, which has here been given short shrift, has apparently employed a form of replication for some time to ensure compatibility of electronic and other parts in order to market them to different companies and applications. Loosely based on this model, the Biological Technologies Office of the US Defense Advanced Research Projects Agency (DARPA) has actually initiated a randomized trial to evaluate the effects of requiring (as a condition of funding) the primary awardees to cooperate and facilitate (sometimes via in person visits or video presentations) independent shadow teams of scientists in the replication and validation of their study results (Raphael, Sheehan, & Vora, 2020). The results of this initiative or its evaluation are not yet available as of this writing, but it is intriguing that the costs that this replication add on typically range between 3% and 8% of the original study's overall budget.

Next

The next chapter looks at two affective views of the replication process based on (a) the reactions of scientists whose original studies have been declared irreproducible and (b) professional (and even public) opinions regarding the career effects thereupon (along with a few hints regarding the best way to respond thereto).

References

Begley, C. G., & Ellis, L. M. (2012). Drug development: raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Boekel, W., Wagenmakers, E.-J., Belay, L., et al. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex, 66*, 115–133.

Camerer, C., Dreber, A., Forsell, E., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour, 2*, 637–644.

Chabris, C., Herbert, B., Benjamin, D., et al. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science, 23*, 1314–1323.

Dreber, A., Pfeiffer, T., Almenberg, J., et al. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*, 15343–15347.

Ebersole, C. R., Athertonb, A. E., Belangeret, A. L., et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental and Social Psychology, 67*, 68–82.

Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biology, 13*, e1002165.

Gertler, P., Baliani, S., & Romero, M. (2018). How to make replication the norm: The publishing system builds in resistance to replication. *Nature, 554*, 417–419.

Harris, R. (2017). *Rigor mortis: How sloppy science creates worthless cures, crushes hope and wastes billions*. New York: Basic Books.

Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6*, 8.

Hughes, P., Marshall, D., Reid, Y., et al. (2007), The costs of using unauthenticated, overpassaged cell lines: How much more data do we need? *Biotechniques, 43*, 575–582.

International Initiative for Impact Evaluation. (2018). http://www.3ieimpact.org/about-us

Kaiser, J. (2016). If you fail to reproduce another scientist's results, this journal wants to know. https://www.sciencemag.org/news/2016/02/if-you-fail-reproduce-another-scientist-s-results-journal-wants-know

Kaiser, J. (2018). Plan to replicate 50 high-impact cancer papers shrinks to just 18. *Science.* https://www.sciencemag.org/news/2018/07/plan-replicate-50-high-impact-cancer-papers-shrinks-just-18

Klein, R., Ratliff, K. A., Vianello, M., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142–152.

Klein, R. A., Vianello, M., Hasselman, F., et al. (2018). Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.

Lorsch, J. R., Collins, F. S., & Lippincott-Schwartz, J. (2014). Cell biology: Fixing problems with cell lines. *Science, 346*, 1452–1453.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives in Psychological Science, 7*, 657–660.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716–1–7.

Osherovich, L. (2011). Hedging against academic risk. *SciBX*, 4.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*, 712–713.

Quan, W., Chen, B., & Shu, F. (2017). Publish or impoverish: An investigation of the monetary reward system of science in China (1999–2016). *Aslib Journal of Information Management, 69*, 1–18.

Raphael, M. P., Sheehan, P. E., & Vora, G. J. (2020). A controlled trial for reproducibility. *Nature, 579*, 190–192.

Reproducibility Project: Cancer Biology. (2012). Brian Nosek (correspondence author representing the Open Science Collaboration). https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology

Schweinsberg, M., Madana, N., Vianello, M., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55–67.

Uhlmann, E. L., Ebersole, C. R., & Chartier, C. R. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science, 14*, 711–733.

Vasilevsky, N. A., Brush, M. H., Paddock, H., et al. (2013). On the reproducibility of science: Unique identification of research resources in the biomedical literature. *PeerJ, 1*, e148.

8

Damage Control upon Learning That One's Study Failed to Replicate

So far we've established that replication is the best (if imperfect) means we have for determining a study's reproducibility. But what happens if one's study, perhaps conducted years ago, fails to replicate and its carefully crafted conclusions are declared to be wrong?

A flippant answer might be that individuals should feel flattered that their study was considered important enough to replicate. Most scientists have not, nor probably ever will be, awarded such a distinction.

And why should colleagues of someone whose study has failed to replicate be surprised since they should know by now that most studies aren't reproducible? But of course we're all human and emotion usually trumps knowledge. And for scientists, their work constitutes a major component of their self-worth.

So the purpose of this brief chapter is to explore the implications for scientists who find themselves on the wrong side of the replication process. And to begin, let's consider two high-profile case studies of investigators who found themselves in this particular situation, the negative aspects of which were exacerbated by the ever increasing dominance of social media—scientific and otherwise.

Case Study 1: John Bargh's Defense of Behavioral Priming

Without getting bogged down in definitions of different types of theories, the more successful a science is, the fewer theories it requires to explain its primary phenomena of interest. Hopefully psychology is an exception here, because a large proportion of its investigators appear to aspire to eventually developing a theory of their very own. And for those not quite ready for this feat, many psychological journal editors encourage investigators to yoke their experiments to the testing of an existing one.

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0009

One of the more popular of these theories is attributed to John Bargh and designated as "behavioral priming"—which might be defined in terms of exposure to one stimulus influencing response to another in the absence of other competing stimuli. Literally hundreds of studies (naturally most of them positive) have been conducted supporting the phenomenon, probably the most famous of which was conducted by John Bargh, Mark Chen, and Lara Burrows (1996), in which participants were asked to create sentences from carefully constructed lists of scrambled words.

The paper itself reported three separate experiments (all positive confirmations of the effect), but the second garnered the most interest (and certainly helped solidify the theory's credence). It consisted of two studies—one a replication of the other—with both asking 30 participants to reconstruct 30 four-word sentences from 30 sets of five words. The participants were randomly assigned to one of two different conditions, one consisting of sets embedded with elderly stereotypic words, the other with neutral words. Following the task they were told to leave the laboratory by taking the elevator at the end of the hall during which a research assistant unobtrusively recorded their walking speed via a stopwatch.

The results of both studies (the original and the self-replication) showed a statistically significant difference between the two conditions. Namely, that the students who had been exposed to the stereotypic aging words walked more slowly than the students who had not—a scientific slam dunk cited more than 5,000 times. Or was it a chimera?

Apparently some major concerns had surfaced regarding the entire concept over the ensuing years, which may have motivated a team of researchers to conduct a replication of the famous study during the magical 2011–2012 scientific window—accompanied, of course by one of the era's seemingly mandatory flashy titles.

Behavioral Priming: It's All in the Mind, But Whose Mind?

Stéphane Doyen, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans (2012)

These investigators attempted to replicate the Bargh team's influential study via the following two experiments. The first was designed as a direct replication of the 1996 study and employed the same intervention

and control tasks (constructing brief sentences from scrambled word sets) followed by the respondents' timed walking speeds.

However, the replication instituted three methodological improvements:

1. The experimenters were *blinded* to group membership (always a requirement in any serious scientific experiment). The original experiment reported that the "experimenter kept himself blind to condition by prepackaging the various scrambled-sentence tasks and picking packets randomly when the participant arrived at the laboratory waiting area" (p. 236).
2. An automated timing mechanism (infrared sensor) was instituted to measure participants' walking speed as they exited the laboratory since the experimenters' use of a stopwatch could theoretically result in biased or inaccurate measurements. And, finally, since Bargh's studies were *severely* underpowered (even for psychology experiments).
3. The replication employed four times as many participants to mitigate that very serious weakness (as explained in Chapter 2).

The bottom line was that despite considerably greater statistical power, the priming effect failed to replicate. But now our investigators got *really* creative and designed their second experiment to ascertain the most plausible reason (inadequate blinding) for Bargh's originally positive findings (as well as their failure to replicate it).

The Blinding Follow-Up Experiment

Fifty undergraduates were randomly assigned to either the priming or non-priming conditions as in the first experiment. Five priming and five control participants were then assigned to each of 10 *experimenters* (i.e., research assistants), five of whom were led to believe that their participants would walk more slowly at the end of the study while the other five experimenters were assured that their participants would walk faster. These "indoctrinations" were an hour long, so obviously a good deal of effort was spent on shaping their expectations in order to ascertain if experimenter beliefs or expectations might have influenced the discrepancy between the original study and its replication. In addition to the electronic timing device, experimenters were asked to also employ a stopwatch as in

the first experiment via the bogus explanation that the investigators had found the infrared sensors to be inaccurate.

The results obtained by the research assistants using stopwatches (which the authors' termed "subjective timing") were as follows:

1. This time around a variation of the Bargh et al. priming effect was replicated in the sense that the primed participants of the five research assistants who had been led to believe that the priming intervention would induce participants to walk more slowly did indeed register significantly slower walking times than their non-primed counterparts.
2. Interestingly, however, the five research assistants who had been led to believe that their primed participants would walk faster actually did record faster walking times than their five control counterparts who had not been fictitiously "sensitized."

And, as the investigators hypothesized, differing results were obtained for the infrared sensors (which it will be recalled involved measuring exactly the same walking behavior for all 50 participants).

1. For the five slow-walking indoctrinated research assistants, their primed participants did walk significantly more slowly than their non-primed participants as in the Bargh et al. study.
2. There was no difference in walking speed between priming and non-priming participants for the fast walking condition.

The investigators therefore concluded, in addition to the fact that despite using a much larger sample they were unable to replicate the Bargh et al. "automatic effect of priming on walking speed," that

> in Experiment 2 we were indeed able to obtain the priming effect on walking speed for both subjective and objective timings. Crucially however, this was *only possible* [emphasis added] by manipulating experimenters' expectations in such a way that they would expect primed participants to walk slower. (p. 6)

John Bargh's Reaction

Although perhaps understandable since much of Bargh's reputation was built on the priming construct (he quoted himself a dozen times in the 1996 study alone), the good professor went absolutely ballistic at the replication results and especially the scientific media's reaction to it. According to Ed Yong (2012), who has written on reproducibility a number of times for *Science* and writes a blog for *Discover Magazine* http://blogs.discovermagazine.com/notrocketscience/failed-replication-bargh-psychology-study-doyen/), Bargh called the authors of the replication "incompetent or ill-informed," defamed the journal *PLoS One* as not receiving "the usual high scientific journal standards of peer review scrutiny" (a damning criticism for which there is absolutely no supporting evidence), and attacked Yong himself for "superficial online science journalism"—all via *his* online blog. (Reactions, incidentally, that Joseph Simmons suggested constituted a textbook case in how *not* to respond to a disconfirming replication.)

The Professional Reaction

Since the Doyen team's failure to replicate Bargh's famous study, the priming construct has understandably lost a considerable amount of its former luster. For example, in the first "Many Labs" replication project (Klein, Ratliff, Vianello, et al., 2014, discussed in the previous chapter), only two replications (of 13) provided no support for their original studies' findings and both of these involved this once popular construct (i.e., flag priming [influencing conservatism] and currency priming [influencing system justification]). And, according to an article in *Nature,* Chivers (2019) maintained that dozens of priming replications did not confirm the effect—perhaps leading Brian Nosek to state concerning the effect that " I don't know a replicable finding. It's not that there isn't one, but I can't name it" (p. 200).

Of course a meta-analysis (Weingarten, Chen, McAdams, et al., 2016) on the topic assessing the word priming effect found a small but statistically significant effect size of 0.33 for 352 effects (not a typo), but that is customary in meta-analyses. However, the average number of participants employed was only 25 per condition, which implied that the average statistical power thereof was 0.20 (which it will be recalled translates to a 20% chance of

obtaining statistical significance if a real effect exists—which in the case of priming an increasing percentage of scientists no longer believe exists).

Case Study 2: The Amy Cuddy Morality Play

What started out as another typically underpowered social psychology experiment involving a short-term, extremely artificial, media-worthy intervention morphed into something that may ultimately have an impact on the discipline comparable to Daryl Bem's psi studies. Perhaps this event served notice that counterintuitive, headline-catching studies could constitute a double-edged sword for their authors by encouraging methodologically oriented critics to mount a social media counterattack—especially if the original investigators appeared to be profiting (financially, publicly, or professionally) by conducting questionable research practice (QRP)-laced, irreproducible science.

What this episode may also herald is an unneeded illustration of the growing use of the internet to espouse opinions and worldviews concerning findings published in traditional peer reviewed journals, a phenomenon that might also reflect a growing tendency for an already broken peer review/editorial propensity to allow political and social biases to influence not just publication decisions, but the hyperbolic language with which those publications are described. (The latter of which is apparently on an upswing; see Vinkers, Tijdink, & Otte, 2015.)

But let's begin by considering the study itself (Carney, Cuddy, & Yap, 2010) whose unimposing title ("Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance") hardly suggested the controversy that would follow. And neither did its abstract portend anything controversial, or at least not until its final sentence.

> Humans and other animals express power through open, expansive postures, and they express powerlessness through closed, contractive postures. But can these postures actually cause power? The results of this study confirmed our prediction that posing in high-power nonverbal displays (as opposed to low-power nonverbal displays) would cause neuroendocrine and behavioral changes for both male and female participants: High-power posers experienced elevations in testosterone,

> decreases in cortisol, and increased feelings of power and tolerance for risk; low-power posers exhibited the opposite pattern. In short, posing in displays of power caused advantaged and adaptive psychological, physiological, and behavioral changes, and these findings suggest that embodiment extends beyond mere thinking and feeling, to physiology and subsequent behavioral choices. *That a person can, by assuming two simple 1-min poses, embody power and instantly become more powerful has real-world, actionable implications* [emphasis added]. (p. 1363)

For those unfamiliar with scientific abstracts, this concluding sentence reflects a completely unacceptable departure from conventional (or at least past) scientific publishing conventions. Had a single word such as "might" been inserted (which a competent journal editor should have insisted upon—even prior 2011–2012 [note the study's publication date of 2010, which is a year before Professor Bem's magnum opus was published]), it is quite possible that the vituperative reactions to the study *might* have been avoided. That and Dr. Cuddy's popularization of her finding.

But alas no such addition (or disclaimer) was added, probably because its authors appeared to belong the school of belief that just about anything to which a $p < 0.05$ can be affixed does translate to "real-world" behavior and scientific meaningfulness. Hence our little morality play, which is eloquently presented in a *New York Times Magazine* piece written by Susan Dominus and aptly titled "When the Revolution Came for Amy Cuddy" (2017, Oct. 18).

Apparently (at least from her critics' perspectives), Dr. Cuddy parlayed her often cited 2010 study (note that she was the second author on the study in question) into two extremely popular YouTube presentations (one of which constituted Ted Talks' second most popular offering with 43 million views and counting), a best-selling book (*Presence*), and myriad paid speaking engagements.

While the study in question was neither her first nor last on the topic, the 2010 article—coupled with her new-found fame—engendered a replication (Ranehill, Dreber, Johannesson, et al., 2015) employing a sample five times as large as the original study's. Unfortunately the original study's positive results for cortisol, testosterone, or risk-taking did not replicate, although a quarter of a point difference was observed on a 4-point rating scale soliciting the degree to which the participants *felt* powerful following the brief set of "power" poses—an effect that was also significant in the original

article but which could be interpreted as primarily an intervention "manipulation check" (i.e., evidence that participants could perceive a difference between the "high-power nonverbal poses" and the "low-power nonverbal poses" but little else).

Naturally Cuddy defended her study as almost any investigator would and continued her belief in the revolutionary societal effects of empowering the un-empowered (e.g., young women, female children, and even black men) via her 1-minute poses. As part of this effort and in response to their critics, she and her original team (note again that Dana Carney was the first author on this paper as well) even accumulated a group of 33 studies involving "the embodied effects of expansive (vs. contractive) nonverbal displays" (Carney, Cuddy, & Yap, 2015), none of which, other than *theirs*, found an effect for testosterone or cortisol. As a group, however, the cumulative results were overwhelmingly positive, which of course is typical of a science which deals in publishing only positive results involving soft, self-reported, reactive outcomes.

Also, in adddtion to appealing to the authority of William James, the authors also presented a list of rebuttals to the Ranehill et al. failure to replicate—one of which was that, for some reason, the latter announced to their participants that their study was designed to test the effects of physical position upon hormones and behavior. (It is unclear why the replicating team did this, although its effect [if any] could have just as easily increased any such difference due to its seeming potential to elicit a demand effect.)

So far, all of this is rather typical of the replication and rebuttal process since no researcher likes to hear or believe that his or her findings (or interpretations thereof) are incorrect. (Recall that even Daryl Bem produced a breathlessly positive meta-analysis of 90 experiments on the *anomalous anticipation of random future events* in response to Galak, LeBoeuf, Nelson, and Simmons's failure to replicate psi.)

But while no comparison between Amy Cuddy and Daryl Bem is intended, Drs. Galak, Nelson, and Simmons (along with Uri Simonsohn) soon became key actors in our drama as well. This is perhaps due to the fact that, prior to the publication of the 33-experiment rejoinder, Dana Carney (again the original first author of both the review and the original power posing study) sent the manuscript along with her version of a p-curve analysis (Simonsohn, Nelson, & Simmons, 2014) performed on these 33 studies to Leif Nelson (who promptly forwarded it on Simmons and Simonsohn).

The *p-curve* is a statistical model designed to ascertain if a related series of *positive* studies' p-values fit an expected distribution, the latter being skewed to the right. Or, in the words of its originators,

> Because only true effects are expected to generate right-skewed *p*-curves—containing more low (.01s) than high (.04s) significant *p* values—only right-skewed *p*-curves are diagnostic of evidential value. (p. 534).

It also doesn't hurt to remember that the p-curve is a statistical model (or diagnostic test) whose utility has not been firmly established. And while it probably is useful for the purposes for which it was designed, its results (like those of all models) do not reflect absolute mathematical certainty. Or, as Stephan Bruns and John Ioannidis (2016) remind us, the exact etiology of aberrant effects (skewness in the case of p-curves) remains "unknown and uncertain."

In any event, at this juncture our story begins to get a bit muddled to the point that no one comes across as completely righteous, heroic, *or* victimized. According to Susan Dominus, Simmons responded that he and Simonsohn had conducted their own p-curve and came to a completely different conclusion from Carney's, whose version they considered to be incorrect, and they suggested that "conceptual points raised before that section [i.e., the "incorrect" p-curve analysis] are useful and contribute to the debate," but, according to Dominus they advised Carney to delete her p-curve and then "everybody wins in that case."

Carney and Cuddy complied by deleting their p-curve but they (especially Amy Cuddy) apparently weren't among the universal winners. Simonsohn and Simmons, after giving the original authors a chance to reply online, then published a decidedly negative blog on their influential *Data Coda* site entitled "Reassessing the Evidence Behind the Most Popular TED Talk" (http://datacolada.org/37), accompanied by a picture of the 1970s television version of *Wonder Woman*. (The latter being a stark reminder of the difference between internet blogs and peer reviewed scientific communication.)

The rest, as the saying goes, is history. Cuddy temporarily became, even more so perhaps than Daryl Bem and John Bargh, the poster child of the reproducibility crisis even though her research was conducted prior to the 2011–2012 enlightenment, as were both Bem's and Bargh's.

Dominus's *New York Times Magazine* article sympathetically detailed the emotional toll inflicted upon Dr. Cuddy, who was portrayed as a brain-injured

survivor who had overcome great obstacles to become a respected social psychologist, even to the point of surviving Andrew Gelman's "dismissive" (Dominus' descriptor) blogs (http://andrewgelman.com/) along with his 2016 *Slate Magazine* article (with Kaiser Fung) critical of her work and the press' role in reporting it. But perhaps the unkindest cut of all came when her friend and first author (Dana Carney) completely disavowed the power pose studies and even recommended that researchers abandon studying power poses in the future.

In response to her colleague's listing of methodological weaknesses buttressing the conclusion that the study was not reproducible, Dr. Cuddy complained that she had not been apprised by Dr. Carney of said problems—which leads one to wonder why Dr. Cuddy had not learned about statistical power and demand characteristics in her graduate training.

So what are we to make of all of this? And why is this tempest in a teapot even worth considering? Perhaps the primary lesson here is that scientists should approach with caution their Facebook, Twitter, and the myriad other platforms that encourage brief, spur-of-the-moment posts. Sitting alone in front of a computer makes it very easy to overstep one's scientific training when responding to something perceived as ridiculous or methodologically offensive.

However. online scientific commentaries and posts are not likely to go away anytime soon and, in the long run, may even turn out to be a powerful disincentive for conducting QRP-laden research. But with that said, perhaps it would be a good idea to sleep on one's more pejorative entries (or persuade a friend or two to serve as one's private peer reviewer). A bit of time tends to moderate our immediate virulent reactions to something we disagree with, which offends our sensibilities, or serves as the subject matter for an overdue blog.

As for Amy Cuddy's situation, it is extremely difficult for some social scientists to avoid formulating their hypotheses and interpreting their results independently of their politico-social orientations—perhaps as difficult as the proverbial camel traversing the eye of a needle. And it may be equally difficult to serve in the dual capacity of scientist *and* entrepreneur while remaining completely unbiased in the interpretation of one's research. Or, even independent of both scenarios, not to feel a decided sense of personal and professional indignation when one's work is subjected to a failed replication and subsequently labeled as not reproducible.

As for articles in the public press dealing with scientific issues, including Susan Dominus's apparently factual entry, it is important for readers to understand that these writers often do not possess a deep understanding of the methodological issues or cultural mores underlying the science they are writing about. And as a result, they may be more prone to allow their personal biases to surface occasionally.

While I have no idea whether any of this applies to Susan Dominus, she did appear to be unusually sympathetic to Amy Cuddy's "plight," to appreciate Joseph Simmons's and Uri Simonsohn's apparent mea culpa that they could have handled their role somewhat differently (which probably would have had no effect on the ultimate outcome), and to not extend any such appreciation to Andrew Gelman, whom she appeared to consider too strident and "dismissive" of Dr. Cuddy.

So while treating Dr. Cuddy's *work* "dismissively" is understandable, it might also be charitable to always include a brief reminder that studies such as this were conducted prior to 2011–2012—not as an excuse but as a potentially mitigating circumstance. If nothing else, such a disclaimer might constitute a subliminal advertisement for the many available strategies for decreasing scientific irreproducibility.

But Is a Failure to Replicate Really All That Serious to a Scientist's Reputation?

If scientists' memories (or attention spans) possess anything resembling those of the general public, the best guess is "not very since it will soon be forgotten." But while no researcher likes to have a cherished finding disparaged, the failure of one's results to replicate is hardly a professional death sentence for anyone's career. But even though opinions and self-reported affects don't amount to a great deal in some sciences, they may have greater importance in the social sciences, hence the following two surveys.

The Ebersole, Axt, and Nosek (2016) Survey

This is the larger survey of the two (4,786 US adults and 313 researchers) with also the more iconic title: "Scientists' Reputations Are Based on

Getting It Right, Not Being Right." Employing brief scenarios in two separate surveys which were then combined, the more germane results of this effort were:

1. In general, investigators who produced replicable but mundane results were preferred to their more exciting counterparts whose results were more questionable.
2. If an investigator's finding replicated, his or her reputation and ability were enhanced (which is rather obvious).
3. If the finding did not replicate, the original investigator's externally perceived ability and ethical behavior decreased somewhat (but definitively more so if he or she criticized the replication study as opposed to agreeing that the original result *might* be wrong or if he or she performed a self-replication in self-defense).

So Simmons might have been on to something when he suggested that John Bargh's online tantrum constituted a textbook case in how *not* to respond to a disconfirming replication. And Ebersole et al. were even kind enough to provide a case study of an exemplary response to a disconfirming replication by Matthew Vees.

> Thank you for the opportunity to submit a rejoinder to LeBel and Campbell's commentary. I have, however, decided not to submit one. While I am certainly dismayed to see the failed attempts to reproduce a published study of mine, I am in agreement with the journal's decision to publish the replication studies in a commentary and believe that such decisions will facilitate the advancement of psychological science and the collaborative pursuit of accurate knowledge. LeBel and Campbell provide a fair and reasonable interpretation of what their findings mean for using this paradigm to study attachment and temperature associations, and I appreciated their willingness to consult me in the development of their replication efforts. (p. 5)

And, as our survey authors opined, a response such as this will, if anything, gain the respect of one's peers.

The Fetterman and Sassenberg Survey (2015)

This survey presented one of four different scenarios to 279 published scientists. The methodology involved in the two scenarios of most interest to us here were described as follows:

> Participants were told to think about a specific finding, of their own, that they were particularly proud of (self-focused). They then read about how an independent lab had conducted a large-scale replication of that finding, but failed to replicate it. Since the replicators were not successful, they tweaked the methods and ran it again. Again, they were unable to find anything. The participants were then told that the replicators published the failed replication and blogged about it. The replicators' conclusion was that the effect was likely not true and probably the result of a Type 1 error. Participants were then told to imagine that they posted on social media or a blog one of the following comments: "in light of the evidence, it looks like I was wrong about the effect" (admission) or "I am not sure about the replication study. I still think the effect is real" (no admission). (p. 4)

(The other two scenarios were practically identical except the description of the replication involved a well-known study published by a prominent researcher to whom the two alternate comments were ascribed.)

The basic results were similar to those of the Ebersole et al. findings. Namely (a) that scientists tend to overestimate the untoward effects of negative replications and (b) these effects are less severe if the original investigators "admit" that they may have been wrong. The authors accordingly speculate that such an admission might repair some of the reputational damage that is projected to occur in these scenarios.

These two surveys produced a plethora of different results that are not discussed here so, as always, those interested in these issues should access the full reports. The Ebersole et al. survey, for example, was able to contrast differences between the general public and scientists with respect to several issues (e.g., researchers were considerably more tolerant of researchers who did not replicate their own research than the general public and were also more appreciative [at least in theory] of those who routinely performed "boring" rather than "exciting" studies). Similarly, Drs. Fetterman and Sassenberg were able to study the relationship between some of their

scenarios and whether or not the respondents were in the reproducibility or business-as-usual research camps.

Both teams transparently mentioned (a) some of the weaknesses of employing scenarios to explain real-life behaviors, (b) the fact that radically different results could be produced by minor tweaks therein, and (c) that the admission of such problems doesn't mitigate their very real potential for themselves creating false-positive results. One potential problem with the interpretation of these surveys, at least in my opinion, is the implication that real-life scientists would be better off admitting that they were (or might have been) wrong when they may have actually believed their original results were correct. (This is ironic in a sense, given Ebersole et al.'s nomination of Matthew Vees's strategy as an exemplary alternative to going into "attack mode" following the failure of one of his study's to replicate since Dr. Vees did not suggest or imply that his initial results might be wrong. Nor did he imply that his replicators were wrong either—which is not necessarily contradictory.)

Of course anyone can quibble about scenario wordings and suggest minor tweaks thereto (which is surely a weakness of scenarios as a scientific tool in general). In truth no one knows whether the results of such studies reflect "real-life" behaviors, reactions, or even if the same results (or interpretation thereof) might change over time. But as imperfect as surveys and scenarios are, the two just discussed at least provide the best assurances available that replication failures, while disappointing, disheartening, and perhaps enraging, are not as bad as they seem to their "victims" at the time. Time may not heal all wounds, but it almost surely will blunt the pain from this one and have very little impact on a career. At least barring fraudulent behavior.

Two Additional Strategies for Speeding the Healing Process Following a Failure to Replicate

The first was proffered by John Ioannidis who, based on meta-analytic evidence, argued that the vast majority of prospective cohort studies indicating a positive relationship between a wide variety of different food intakes and mortality risk probably constituted false positive results.

> The nutritional epidemiology community includes superb scientists. The best of them should take ownership of this reform process. They can further

> lead by example (e.g., by correcting their own articles that have misleading claims). Such corrections would herald high scientific standards and public responsibility. (Ioannidis, 2018, p. 970)

At first glance this plea may seem somewhat Pollyannaish and destined to be ignored, but with a bit of thought it may be excellent advice for preempting a negative replication of one's work regardless of discipline. Many investigators probably suspect at least one of their published studies to represent a false-positive result. So, rather than hunkering down in the dark in the hope that some reproducibility miscreant with nothing better to do will decide to replicate said study, why not replicate it oneself (perhaps involving one's original co-authors)?

If the study replicates, great! If it does not, it will represent another publication and label the original author as a principled, 21st-century scientist.

The second strategy was offered by Jarrod Hadfield (2015), an evolutionary biologist, who proposed a strategy for implementing Ioannidis's plea. Professor Hadfield suggested that since investigators often "continue to collect data on their study systems that could be used to validate previous findings. [Thus a]llowing authors to publish short (two paragraph) addenda to their original publications would lower the costs of writing and submitting replication studies, and over time these addenda may reduce the stigma associated with publishing false positives and increase transparency."

He was also reported by Forstmeier, Wagenmakers, and Parker (2017) as somewhat iconoclastically suggesting

> that researchers running long-term studies [to which longitudinal studies stretching over decades, such as the Framingham Heart Study, would qualify] should publish addenda to their previous publications, declaring in a one-page publication that their original finding did or did not hold up in the data of the following years (after the publication), and comparing the effect sizes between the original data and the newer data. This would be a quick way of producing another publication, and it would be enormously helpful for the scientific field. This may also relax the feeling of stigma when something does not hold up to future evaluation. Admitting a failure to replicate could actually be perceived as a signal of a researcher's integrity and be praised as a contribution to the scientific community." (p. 1960)

Which Provides a Conveniently Lead-In to Chapter 9

To this point, some derivation of the word "publish" has been mentioned more than 200 times in several different contexts—many suggesting various biases due to the various actors involved in the process or inherent to the process itself. It seems natural, therefore, that these factors should be examined in a bit more detail from the perspectives of (a) identifying some of their more salient characteristics that impede the replication (and the scientific) process, (b) facilitating the prevalence of false-positive results in the scientific literature, and, of course, (c) examining a number of suggestions tendered to ameliorate these problems. All of which constitute the subject matter of Chapter 9.

References

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype-activation on action. *Journal of Personality and Social Psychology, 71*, 230–244.

Bruns, S. B., & Ioannidis, J. P. A. (2016). P-curve and p-hacking in observational research. *PLoS One 11*, e0149144.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21*, 1363–1368.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2015). Review and summary of research on the embodied effects of expansive (vs. contractive) nonverbal displays. *Psychological Science, 26*, 657–663.

Chivers, T. (2019). What's next for psychology's embattled field of social priming. *Nature, 576*, 200–202.

Dominus, S. (2017). When the revolution came for Amy Cuddy. *New York Times Magazine*, Oct. 18.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*, e29081.

Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2016). Scientists' reputations are based on getting it right, not being right. *PLoS Biology, 14*, e1002460.

Fetterman, A. K., & Sassenberg, K. (2015). The reputational consequences of failed replications and wrongness admission among scientists. *PLoS One*, *10*, e0143723.

Forstmeier, F., Wagenmakers, E-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings: a practical guide. *Biological Reviews, 92*, 1941–1968.

Gelman, A., & Fung, K. (2016). The power of the "power pose": Amy Cuddy's famous finding is the latest example of scientific overreach. https://slate.com/technology/2016/01/amy-cuddys-power-pose-research-is-the-latest-example-of-scientific-overreach.html

Hadfield, J. (2015). There's madness in our methods: Improving inference in ecology and evolution. https://methodsblog.com/2015/11/26/madness-in-our-methods/

Ioannidis, J. P. A. (2018). The challenge of reforming nutritional epidemiologic research. *Journal of the American Medical Association, 320*, 969–970.

Klein, R., Ratliff, K. A., Vianello, M., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142–152.

Ranehill, E., Dreber, A., Johannesson, M., et al. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26*, 653–656.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *British Medical Journal, 351*, h6467.

Weingarten, E., Chen, Q., McAdams, M., et al. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally-presented words. *Psychological Bulletin, 142*, 472–497.

Yong, E. (2012). Bad copy. *Nature, 485*, 298–300.

PART III

STRATEGIES FOR INCREASING THE REPRODUCIBILITY OF PUBLISHED SCIENTIFIC RESULTS

9

Publishing Issues and Their Impact on Reproducibility

The vast majority of the studies or opinions cited to this point have been published in peer reviewed journals. Historically, scientific publishing has taken many forms and is an evolving process. In past centuries, books were a primary means of communicating new findings, which is how James Lind announced his iconic discovery regarding the treatment and prevention of scurvy in 1753 (*Treatise of the Scurvy*). Which, of course, was ignored, thereby delaying a cure being adopted for almost half a century.

Gradually scientific journals began to proliferate, and as many as a thousand were created during Lind's century alone. Peer reviewed journals now constitute the primary medium for formally presenting new findings to the scientific community, acting as repositories of past findings, and forming the foundation on which new knowledge is built.

So, obviously, the issue of scientific reproducibility cannot be considered in any depth without examining the publishing process itself. Especially since over half of published studies may be incorrect.

Publishing as Reinforcement

The adage of "publish or perish" is a time-honored academic expression, often rued as a professional curse and more recently as a contributor of false-positive results. Nevertheless, peer reviewed publications are the firmly entrenched coin of the scientific realm as surely as money is for investment banking.

Unfortunately, the number of publications to which a scientist's name is attached has sometimes becomes a compulsive end in and of itself, especially when it governs how some scientists, institutions, and funding agencies judge themselves and others. This is a compulsion reflected by number 10 of our inane scientific policies (Chapter 3) in which a surprising number of

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0010

individuals publish one paper every 5 days (Ioannidis, Klavans, & Boyack, 2018)—some of which are never cited and probably never even read by their co-authors.

However, with all of its deficiencies as a metric of scientific accomplishment, publishing is an essential component of the scientific enterprise. And, in an era in which traditional scientific book publishing appears to be in decline, academic journal publishing has evolved relatively quickly into a multibillion-dollar industry.

Whether most scientists approve of the direction in which publishing is moving is unknown and probably irrelevant. Certainly nothing is likely to change the practice of publishing one's labors as a reinforcement of behavior or the narcotic-like high that occurs when investigators receive the news that a paper has been approved for publication. (Of course a "paper," while still in use as an antiquated synonym for a research report, is now digitally produced and most commonly read online or downloaded and read as a pdf—all sans the use of "paper.")

So let's take a quick look at the publishing process through the lens of scientific reproducibility. But first, a few facts that are quite relevant for that purpose.

The Scope of the Academic Publishing Enterprise

Every 2 years the National Science Foundation (NSF) produces a report entitled *Science and Technology Indicators*, which is designed to encapsulate the scientific institution as a whole. As of this writing, the current version (National Science Board, 2018) is an impressive and exhaustive volume running to well over a thousand pages, dealing with the national and international scientific enterprise as a whole including, among other things, the number of science, technology, engineering, and mathematics (STEM) personnel, PhDs awarded, patents granted, and, of course, scientific reports published.

Most relevant to our purposes here are its data on the last indicator, which, for 2016 alone (the most recent year covered in the 2018 publication) totaled a mind-boggling 2,295,608 published entries, a figure that has been increasing at a rate of 3.9% *per year* for the past 10 years. And while a 46.5% decadal increase in the number of publications is astonishing and bordering on absurd, some scholars consider the number of publications as a positive,

planet-wide indicator of scientific progress. (It should be mentioned that Björk, Roos, and Lauri [2009] take issue with this projection and argue that 1,275,000 is a more accurate figure, but both estimates are mind-boggling.)

As shown in Table 9.1 (abstracted from table 5-6 of the original NSF report) only 18% of these publications emanated from the United States, and the majority (54%) of the 2 million-plus total involved disciplines not or only spottily covered in this book (e.g., engineering, the physical sciences, mathematics, computer sciences, agriculture). However, this still leaves quite a bit of publishing activity in almost all recognized disciplines.

Psychology alone contributed more than 39,000 publications, with the other social sciences accounting for more than three times that number. How many of these included p-values are unknown, but, given the modeling studies discussed earlier, psychology alone undoubtedly produces thousands of false-positive results per year with the other social sciences adding tens of thousands more based on their combined 2016 output of more than 120,000 publications. So let's hope that Ioannidis (2005) was wrong about *most* disciplines being affected by his pessimistic modeling result. Otherwise, we're

Table 9.1 Numbers of scientific publications in 2016

Discipline	World	United States	European Union	China
Total publications, all disciplines	**2,295,608**	**408,985**	**613,774**	**428,165**
Engineering	422,392	50,305	89,611	123,162
Astronomy	13,774	3,272	5,524	1,278
Chemistry	181,353	20,858	41,123	53,271
Physics	199,718	27,402	50,943	42,190
Geosciences	130,850	20,449	33,758	30,258
Mathematics	52,799	8,180	15,958	8,523
Computer sciences	190,535	26,175	52,785	37,076
Agricultural sciences	50,503	4,908	12,275	9,376
Biological sciences	351,228	73,208	92,066	59,663
Medical sciences	507,329	119,833	149,761	56,680
Other life sciences	27,547	9,816	7,979	852
Psychology	39,025	14,314	12,889	1,278
Social sciences	121,667	29,447	49,102	4,262

talking about hundreds of thousands (perhaps more than a million) of false-positive scientific reports being produced each year worldwide.

The Current Publication Model

While everyone reading this book probably already knows the basics of the subscription-based 20th-century journal publication process, let's review a few of its main components in order to examine how it has changed in this increasingly digital age (and may even effectively disappear altogether in a few decades), what some of its limitations are, and some rather radical suggestions for its improvement (or least alteration). If readers will forgive a brief self-indulgence, I will begin this attempt by detailing some of my experiences as an editor-in-chief as an example of these relatively rapid changes.

In 1978, an unfortunately now-deceased faculty colleague and I decided to establish a peer reviewed program evaluation journal dedicated to the health sciences. In those pre-email days all correspondence was done by mail (usually on an IBM Selectric typewriter somewhere) hence communication with reviewers, the editorial board, and authors involved enumerable trips back and forth to the post office. After a few issues in which we two academics served as publishers, printers, mailers, promoters, editors, solicitors of articles (at least in the journal's early days), selectors (and solicitors) of peer reviewers, and collators of the latter's often disparate and contradictory reviews, all while maintaining additional back-and-forth correspondence with authors regarding revisions or reasons for rejections ad nauseam, we were quite happy to sell the journal for a whopping $6,000 each to Sage Publications (making it that publisher's *second* peer reviewed, professional journal).

Four decades later, Sage now publishes more than a thousand journals and is dwarfed by other journal publishers such as Elsevier and Wolters Kluwer. All communications now employ the internet rather than the postal service, all of which pretty much encompasses the evolution of academic publishing prior to exclusively online publishers—that and the price changes, with the average library subscription for a scientific journal now being around $3,000 per year plus the substantial (sometimes exorbitant) publishing fees levied by many journals on authors for the privilege of typesetting their labors and distributing them to (primarily) institutional subscribers. (The actual number

of legitimate [i.e., non-predatory] peer reviewed scientific journals is actually not known but it is undoubtedly well over 25,000.)

The Peer Review Process

The hallmark of scientific publishing involves the use of other scientists (i.e., peers) to review their colleagues' work in order to determine its significance, quality, and whether or not it should be published. Practices vary for the peer review process, but typically once a manuscript is submitted for publication someone gives the manuscript a quick read (or perusal) to ensure its appropriateness for the journal in question.

Who this "someone" is varies with the size of the journal and the resources available to it, but typically three peer reviewers are solicited who hopefully are conversant with the topic area in question. Assuming that these individuals comply with the request, the manuscript is forwarded to them accompanied by instructions, a checklist of some sort (hopefully), a time table for their response (often not met), and a request for comments accompanied by a bottom-line decision regarding publication or rejection. Unfortunately, after all of this time and effort, it is not unusual for three disparate decisions emanating from the process such as one reviewer decreeing "publish with minor revisions," another suggesting that the authors make the requested revisions followed by another round of reviews, and the third rejecting the paper out of hand.

Fortunately for everyone concerned, the entire process is now conducted online, but, as described later, the process can take an inordinate of time and is becoming increasingly unwieldy as the number of journals (and therefore submitted papers) proliferate. The present output of 2,295,608 annual papers, for example, could theoretically require 6,886,824 peer reviews if every submission was published and only three reviewers were used. However, if the average journal accepts only half of its reviewed manuscripts (a large proportion typically accept less than 10% of submissions), the peer review burden jumps to north of 10 million reviews, not counting resubmitted rejections to other journals.

Thus, the bottom line of even online publication is that the peer review process has become impossibly overwhelmed and the quality of those reviews has become extremely suspect since many lower tiered journals are forced to take whoever they can get (which often equates to reviewers who

have little or no expertise regarding what they are reviewing). And things are likely to get worse as more and more scientists attempt to publish ever greater numbers of studies.

However, to be fair, the peer review process was always a few orders of magnitude short of perfection. Until a few years ago, no one had come up with a better solution, and it did seem to work reasonably well, especially for the larger and more prestigious journals.

True, recently artificial intelligence aids have been developed that may *slightly* facilitate handling the increasing glut of manuscripts to be reviewed. These platforms can now perform cursory supplemental tasks, as succinctly described by Heaven (2018) including checking for problematic statistical or procedural anomalies (e.g., ScholarOne, StatReviewer), summarizing the actual subject matter of an article rather than relying on a quick perusal of its abstract (which one sage describes as "what authors come up with five minutes before submission" [courtesy of a marketing director in the Heaven paper]), and employing automated plagiarism checks. But regardless of the sophistication achieved by these innovations, the heavy lifting will remain with scientists as long as the current system exists.

Like every other scientific topic, a substantial literature has grown up around the shortcomings of peer review. While this literature can't be done justice here, one of the most thorough and insightful discussions of the entire publishing process (accompanied by potential solutions) must surely be Brian Nosek and Yoav Bar-Anan's (2012) essay entitled "Scientific Utopia: I. Opening Scientific Communication."

However, prior to considering some of these quite prescient suggestions, let's consider the following either amusing or alarming study, depending on one's perspective.

Who's Afraid of Peer Review?

John Bohannon (2013)

Actually, this paper is as much about the dark side of predatory open-access journals as it is about the peer review process but let's continue to focus our attention upon the latter. Financed by *Science* magazine (which perhaps not coincidentally is definitely not an open-access journal and may have had a conflict of interest here), John Bohannon wrote a

completely fake article under a fictitious name from a fictitious university ballyhooing the promise of an also fake cancer drug. (Several versions of the paper were prepared, and the language was purposefully distorted by translating it into French via Google Translate and then back again into English. However, all versions involved a fictitious drug that had not gone to trial, and the article was purposefully filled with so many obvious flaws that [according to the author] no competent reviewer would recommend its acceptance.)

The article was then submitted to 304 open-access journals and was accepted by more than half of them, all with no notice of the study's fatal methodological flaws. As one example, the *Journal of Natural Pharmaceuticals* (published by an Indian company which owned 270 online journals at the time but has since been bought by Wolters Kluwer, a multinational Netherland publishing behemoth with annual revenues of nearly $5 billion) accepted the article in 51 days with only minor formatting changes requested. Nothing was mentioned concerning the study flaws.

For the exercise as a whole, the author reported that

> The paper was accepted by journals hosted by industry titans Sage and Elsevier. The paper was accepted by journals published by prestigious academic institutions such as Kobe University in Japan. It was accepted by scholarly society journals. It was even accepted by journals for which the paper's topic was utterly inappropriate, such as the *Journal of Experimental & Clinical Assisted Reproduction*. (p. 61)

Incredibly, only *PLoS One* (the flagship journal of the Public Library of Science and much maligned by John Bargh) rejected the paper on methodological grounds. One of Sage's journals (*Journal of International Medical Research*) accompanied its acceptance letter with a bill for $3,100. (For many such acceptances, Bohannon sent an email withdrawing the paper due to an "embarrassing mistake.")

Perhaps the main conclusion that can be drawn from this iconoclastic "survey" is that everything in science, as in all other human pursuits, can be (and often is) gamed. Other examples designed to expose the problems

associated with peer review abound. For example, MIT students used SCIgen, a computer program that automatically generates gobbledygook papers, to submit papers that somehow got through the peer review process; a recent group of bizarre fake articles and authors were published in small peer review journals with titles such as "Human Reactions to Rape Culture and Queer Performativity at Urban Dog Parks in Portland, Oregon" (Wilson [Retracted, 2018] published in *Gender, Place, & Culture: A Feminist Geography Journal*) and Baldwin [Retracted, 2020] "Who Are They to Judge?" and "Overcoming Anthropometry Through Fat Bodybuilding" (published in the journal *Fat Studies* [yes, this and the feminist geography journal are actual journals]); and, of course, the iconic hoax by physicist Alan D. Sokal, who published a completely nonsensical article allegedly linking the then post-modernism fad with quantum physics in the non-peer reviewed *Social Text* (1996) (https://en.wikipedia.org/wiki/Alan_Sokal).

One problem with both the publishing and peer review processes lies in the number of publishing outlets available to even borderline scientists, which means that just about anything can be published with enough perseverance (and money). Another lies in the mandate that journals (especially subscription-based ones) are typically required to publish a given number of articles every quarter, month, or in some cases week. This dilutes the quality of published research as one goes down the scientific food chain and may even encourage fraudulent activity, such as the practice of authors nominating actual scientists, accompanied by fake email addresses (opened solely for that purpose and sometimes supplied by for-profit companies dedicated to supporting the process). Springer, for example (the publisher of *Nature* along with more than 2,900 other journals and 250,000 books), was forced to retract 107 papers from *Tumor Biology* published between 2010 and 2016 due to fake peer reviews. Similarly, Sage Publications was forced to retract 60 papers (almost all with the same author) due to a compromised peer review system. And these are known and relatively easily identified cases. We'll probably never know the true extent of these problems.

On a lighter note, Ferguson, Marcus, and Oransky (2014) provide several interesting and somewhat amusing examples of peer review fraud along with potential solutions. (Cat Ferguson, Adam Marcus and Ivan Oransky are the staff writer and two co-founders, respectively, of *Retraction Watch*, an extremely important organization designed to track retracted papers and whose website should be accessed regularly by all scientists interested in reproducibility.)

Examples supplied by the group include:

1. The author asks to exclude some reviewers, then provides a list of almost every scientist in the field.
2. The author recommends reviewers who are strangely difficult to find online.
3. The author provides gmail, Yahoo, or other free e-mail addresses to contact suggested reviewers, rather than email addresses from an academic institution.
4. Within hours of being requested, the reviews come back. They are glowing.
5. Even reviewer number three likes the paper (p. 481). [In my experience three-for-three uncritically positive reviews are relatively uncommon.]

Predatory (Fake) Journals

As mentioned, another downside of the publishing process involves the proliferation of predatory journals that will publish just about anything for "a few dollars more." We can blame this one on the inexorable movement toward open access, online publishing which ironically was originally motivated by the most idealistic of goals: making scientific knowledge open to everyone.

Regardless of who or what is to blame (if anyone or anything other than the perpetrators themselves), most of these journals appear to be located in India and China but often pretend to be located in the United States—although the latter contributes its share of home-grown outlets as well. Besides scamming gullible, desperate, and/or inexperienced investigators, predatory journals' characteristics include no true peer review system (just about every submission is accepted), no subscription base, a title that neither the submitting investigators nor their colleagues have ever heard of, websites replete with misinformation, and impact factors approaching or including zero.

Declan Butler (2013) provides an excellent overview of this problem in an article entitled "The Dark Side of Publishing," as well as suggesting several strategies for identifying suspect journals.

- Check that the publisher provides full, verifiable contact information, including an address, on the journal site. Be cautious of those that provide only web contact forms.

- Check that a journal's editorial board lists recognized experts with full affiliations. Contact some of them and ask about their experience with the journal or publisher since sometimes these journals simply list prominent scientists without their knowledge.
- Check that the journal prominently displays its policy for author fees.
- Be wary of email invitations to submit to journals or to become an editorial board member. [This one is tricky since legitimate journals sometimes use email correspondence to solicit manuscripts for a special issue or to contact potential board members based on recommendations from other scientists.]
- Read some of the journal's published articles and assess their quality. Contact past authors to ask about their experience. [This, too, has a downside since some of the authors may be quite proud of the fact that their articles were accepted with absolutely no required revisions.]
- Check that a journal's peer review process is clearly described, and try to confirm that a claimed impact factor is correct.
- Find out whether the journal is a member of an industry association that vets its members, such as the Directory of Open Access Journals (www.doaj.org) or the Open Access Scholarly Publishers Association (www.oaspa.org).
- Use common sense, as you would when shopping online: if something looks fishy, proceed with caution. (p. 435)

But Is an Actual Paradigmatic Publishing Change Actually Needed?

Certainly just about every reproducibility methodologist mentioned here would agree that some changes are desperately needed, although the "paradigmatic" adjective might appear unnecessarily radical to some. However, suggested change of any sort often appears radical when first proposed, so let's consider a few of these suggested changes here—some radical, some paradigmatic, and some simply sensible.

And who better to present these suggestions than Brian Nosek and one of his like-minded colleagues via the following forward-thinking essay.

Scientific Utopia I. Opening Scientific Communication

Brian Nosek and Yoav Bar-Anan (2012)

This long, comprehensive article clearly delineates the problems bedeviling publishing in peer reviewed scientific journals followed by proposed solutions for each. Ideally it should be read in its entirety by anyone interested in reforming the current system, but, for present purposes, what follows is an encapsulated version of its authors' vision of both some of the problems with the current system and their potential solutions.

First the problems:

1. The long lag between submission of an article and its appearance in print (perhaps averaging close to 2 years);
2. The astonishing subscription costs to university libraries;
3. The inaccessibility of scientific articles to anyone not affiliated with a subscribing institution;
4. *The odd arrangement of scientists turning over ownership of their articles (for which they are not paid) to publishers to print their work in one of the publisher's journals* (emphasis is added since the arrangement truly is odd and because scientists or their institutions typically have to *pay* these publishers for the honor of accepting said ownership.);
5. The myriad problems with the peer review system;
6. The static nature of a journal article that, once printed, stays printed and remains unchanged while science itself is a fluid and ever-changing process (this is also quite odd since, among other absurdities, author-initiated errata are often published in later journal issues instead of being corrected in the original digital version of the article. And to add insult to injury, authors are often charged for their errata or their requested withdrawal [i.e., retraction] of an article);
7. And, finally, the limited amount of space in printed journals, making it only possible to communicate what the investigators and reviewers consider essential information (but almost never enough information to allow a study to be replicated).

Needless to say Drs. Nosek and Ban-Anan have potential solutions for each of these shortcomings.

Fully embracing digital communication and completely abandoning journal issues and paper copies. Of the 25,000 or so scientific journals, the majority are online only. (Actually no one knows for sure just how many journals exist, although Björk et al., 2009, reported that there were 23,750 in 2006, which was over a decade ago.) Today almost everyone reads journal articles online or downloads pdfs to read at their leisure. Paper copies of all but the most widely read journals are disappearing from academic libraries so part of this suggestion is well on its way to being implemented. As for journal issues, little would be lost and precious time gained if the increasingly popular practice of making the final version of papers available to subscribers online in advance of the completed issue were to simply replace the issue system itself. The PLoS model, for one, already does this by making papers freely available as soon as they are accepted, suitably revised, and copyedited.

Going to a totally open access model in which all scientists (whether they work at universities, for private industry, or in their own basements) can access *everything* without costs. Of course the obvious problem with this is that *someone* has to pay the costs of copyediting and other tasks, but going to an exclusively digital model (and possibly, ultimately, a nonprofit one) should greatly reduce costs. Nosek and Bar-Anan suggest that most of these costs should be borne by scientists, their funding agencies, or their institutions (perhaps augmented by advertisers).

Again the PLoS open-access, purely digital model is presented as one example of how this transition could be made. Another example is the National Institutes of Health (NIH)'s PubMed Central (http://www.ncbi.nlm.nin.gov/pmc/), which attempts to ensure open access to all published reports of NIH-funded research. The greatest barriers to the movement itself are individual scientists, and a number of suggestions are made by the authors to encourage these scientists to publish their work in open-access journals. The major disadvantage resides in the inequitable difficulty that unfunded or underfunded scientists will have in meeting publication costs (which are considerable but are often also charged by subscription outlets) although some open-access journals theoretically reduce or even waive these fee if an investigator has no dedicated funding for this purpose.

Publishing prior to peer review. Citing the often absurd lag between study completion and publication, the authors suggest that "Authors prepare their manuscripts and decide themselves when it is published by submitting it to a repository. The repository manages copyediting and makes the articles available publicly" (p. 231).

Examples of existing mechanisms through which this is already occurring are provided, the most notable being the previously mentioned decades-old and quite successful arXiv preprint repository (https://en.wikipedia.org/wiki/ArXiv), followed by a growing group of siblings (some of which allow reviewer comments that can be almost as useful as peer reviews to investigators). An important subsidiary benefit for scientists is avoiding the necessity of contending with journal- or reviewer-initiated publication bias.

In this model, preprints of manuscripts are posted without peer review, and under this system, the number of submissions for arXiv alone increased by more than 10,000 *per month* by 2016. Most of the submissions are probably published later in conventional outlets, but, published or not, the repository process has a number of advantages including

1. Allowing papers to be revised at the scientists' pleasure, thus becoming a sort of living document (relatedly, it might be wise for even preprint registries to require some disciplinary-appropriate version of Simmons, Nelson, and Simonshohn's "21-word solution");
2. Aiding scientists in keeping up with advances in the field much more quickly than waiting on the snail-paced conventional publishing system;
3. Establishing the *priority* of a discovery since the preprinted paper can be posted almost immediately after the discovery is made instead of waiting months or years in the conventional journal system (this would also reduce investigator paranoia that someone might steal his or her discovery or idea for one);
4. Greatly reducing publication bias since (a) no one can reject the preprint because it wasn't accompanied by a favorable p-value and (b) the process increases investigator incentives to deposit their negative studies in a public repository rather than in their unrewarded file drawers;
5. Preventing costly redundancies by allowing other investigators to build on work that they would otherwise have to perform

themselves (of course, replications would remain an important scientific activity, just a more selective one);
6. And, in some cases (where professional comments are permitted), providing a valuable form of peer review preceding formal journal submissions of manuscripts.

However, there is a moral here, and it is that even extremely promising innovations such as preprint repositories that appear to be available free of cost must be financed somehow. And, as an ironical example of this, the Center for Open Science (COS) repository (ironic since Brian Nosek serves as its executive director) has announced that, as of 2019, it will begin charging fees to the 26 other organizations' repositories that it hosts since its projected costs for the year 2020 of $260,000 can no longer be covered by grant funding. And, as only one example, the INA-Rxiv alone, which now receives more than 6,000 submissions per year, will be faced with $25,000 in annual fees and accordingly has decided to leave the COS repository (Mallapaty, 2020).

Making peer review independent of the journal system. Here the authors really hit their stride by suggesting the creation of general or generic peer review systems independent of the journals themselves. In this system "instead of submitting a manuscript for review by a particular journal with a particular level of prestige, authors submit to a review service for peer review . . . and journals become not the publisher of articles but their 'promotors' " (p. 232).

This process, the authors argue, would free journals from the peer review process and prevent investigators from the necessity of going through the entire exercise each time they submit a paper following a rejection. Both the graded results of the reviews and the manuscript itself would be available online, and journal editors could then sort through reviewed articles according to their own quality metrics and choose which they wished to publish. More controversially, the *authors also suggest that there would be no reason why the same article couldn't be published by multiple journals*. (How this latter rather odd strategy would play out is not at all clear, but it is a creative possibility and the suggestion of employing a single peer review process rather than forcing each journal to constitute its own, while not unique to these authors, is in my opinion as a former editor-in-chief absolutely *brilliant*.)

Publishing peer reviews. Making a case for the important (and largely unrewarded) contribution made by peer reviewers, the authors suggest that peer reviews not only be published but also not be anonymous unless a reviewer so requests. This way reviewers could receive credit for their scientific contributions since, as the authors note, some individuals may not have the inclination, opportunity, or talent for *conducting* science but may excel at evaluating it, identifying experimental confounds, or suggesting alternative explanations for findings. In this scenario, reviewers' vitas could correspondingly document this activity.

It might also be possible for "official" peer reviewers to be evaluated on both the quality of their reviews and their evaluative tendencies. (Many journal editors presently do this informally since some reviewers inevitably reject or accept every manuscript they receive.) Listed advantages of these suggestions include the avoidance of "quid pro quo positive reviewing among friends" as well as the retaliatory anonymous comments by someone whose work has been contradicted, not cited, or found not to be replicable by the study under review.

Continuous, open peer review. Peer reviews are not perfect and even salutary ones can change over time, as occurs when a critical confound is identified by someone following initial review and publication or a finding initially reviewed with disinterest is later found to be of much greater import. The authors therefore suggest that the peer review process be allowed to continue over time, much as book reviews or product evaluations do on Amazon.com. To avoid politically motivated reviews by nonscientists, a filter could be employed, such as the requirement of an academic appointment or membership in a professional organization, in order to post reviews. (This latter suggestion could be buttressed by the creation of an interprofessional organization—or special section within current professional organizations—devoted to the peer review process.)

Naturally, there are other insightful suggestions for reforming the publishing and peer review process not mentioned in this seminal article, but this one certainly constitutes the most comprehensive discussion of possible reforms of which I am aware. One such unmentioned innovation that has witnessed a degree of implementation follows. *"As is" peer review*. Eric Tsang and Bruno Frey (2007) have proposed an "as is" review process that they argue should, among other things, shorten the review process for everyone

involved and reduce "the extent of intellectual prostitution" occurring when authors accede to sometimes inappropriate revisions by reviewers less knowledgeable than themselves.

The authors also note that while peer reviews in management journals (their discipline) were originally a page or less in length, reviews had mutated to eight or more single-spaced pages by the middle of the 1980s, followed by editorial requirements that authors supply point-by-point responses thereto. All of which were often followed by more reviews and more responses, the totality of which sometimes exceeded the length of the original article.

Tsang and Frey accordingly proposed a process (some variant of which has subsequently been implemented by a number of journals) that would involve the reviewers' providing feedback as currently practiced but then restricting their bottom-line decision to only an accept or reject option (not the dreaded "resubmission for further review" based on minor or major revisions). Then, in the authors' words,

> Based on the referees' recommendations, and his or her own reading of the manuscript, the editor makes the decision to accept or reject the manuscript. If the editor accepts the manuscript (subject to normal copy editing), he or she will inform the authors accordingly, enclosing the editorial comments and comments made by the referees. It is up to the authors to decide whether, and to what extent, they would like to incorporate these comments when they work on their revision for eventual publication. As a condition of acceptance, the authors are required to write a point-by-point response to the comments. If they refuse to accept a comment, they have to clearly state the reasons. The editor will pass on the response to the referees. In sum, the fate of a submitted manuscript is determined by one round of review, and authors of an accepted manuscript are required to make one round of revision. (pp. 11–12)

There are possible variations on this, as well as for all the proposals tendered in this chapter for reforming the publication and peer review process. All have their advantages, disadvantages, and potential pitfalls, but something has to be changed in this arena if we are to ever substantively improve the reproducibility of empirical research.

Other Publishing Issues Impacting Reproducibility

Retractions of Published Results

So far we haven't discussed the retraction of erroneous research findings in the detail that the process deserves since its proper application has the potential of reducing the prevalence of irreproducible results. However, although the handling of retractions is an important component of the publication process, they appear to have often been treated as an unwelcomed stepchild.

In my opinion (and one which I feel confident is shared by Adam Marcus and Ivan Oransky of Retraction Watch [https://retractionwatch.com/]), many journal editors and their publishers are several steps beyond passive aggression when it comes to handling retractions. In addition to the already mentioned tendency for journals to require publication fees from authors, many charge an additional fee for anyone who wishes to retract a study, correct a mistake, or even alert readers to an egregious error not reported by the original authors.

A ridiculous example of this reluctance on the part of journals to acknowledge the existence of such errors is provided by Allison, Brown, George, and Kaiser (2016), who recount their experiences in alerting journals to published *errors* in papers they were reviewing for other purposes. They soon became disenchanted with the process, given that "Some journals that acknowledged mistakes required a substantial fee to publish our letters: we were asked to spend our research dollars on correcting other people's errors" (p. 28).

Of course, some retractions on the part of investigators reflect innocent errors or oversights, but apparently most do not. Fang, Steen, and Casadevall (2012), for example, in examining 2,047 retractions indexed in PubMed as of 2012 found that 67.4% were due to misconduct, fraud, or suspected fraud. And what is even more problematic, according to the Retraction Watch website, some articles are actually cited more frequently *after* they are retracted than before.

Exactly how problems such as this can be rectified is not immediately clear, over and above Drs. Marcus and Oransky's continuing Herculean efforts with Retraction Watch. One possibility that could potentially put a dent in the problem, however, is to send a corrective email to any investigator citing a retracted article's published results, perhaps even suggesting that he or she retract the citation.

Should Scientists Publish Less Rather Than More?

Undoubtedly some should, but who is to decide who and how much? Many of the Utopia I article's recommendations would probably result in a significant increase in per scientist published outputs, and whether or not this is desirable is open for debate.

Brian Martinson (2017) makes a persuasive case for some of overpublication's undesirable consequences, and few scientists would probably disagree with it (at least in private).

> The purpose of authorship has shifted. Once, its primary role was to share knowledge. *Now it is to get a publication* [emphasis added]—"pubcoin: if you will. Authorship has become a valuable commodity. And as with all valuable commodities, it is bought, sold, traded and stolen. Marketplaces allow unscrupulous researchers to purchase authorship on a paper they had nothing to do with, or even to commission a paper on the topic of their choice. "Predatory publishers" strive to collect fees without ensuring quality. (p. 202)

However, many fewer would agree with his solution of giving scientists a "lifetime word limit" which Martinson himself freely admits might have a number of negative consequences. Alternately, Leif Nelson, Joseph Simmons, and Uri Simonsohn (2012), in responding to the Nosek and Bar-Anan "Utopia" paper, floated a one paper per year alternative (a solution not without its own advantages and drawbacks, but which is probably as equally unlikely to be implemented as Martinson's suggestion).

More importantly, however, the Nelson et al. response also makes a strong case against relaxing the publication process to the point where just about *everyone* can publish just about *anything*—a *possible* outcome if some of the more radical Nosek and Bar-Anan proposals were to be implemented.

> Bad papers are easy to write, but in the current system they are at least *somewhat* [emphasis added] difficult to publish. When we make it easier to publish papers, we do not introduce good papers into the market (those are already going to be out there); we introduce disproportionately more bad papers. (p. 292)

And then, of course, there is always John Ioannidis and colleagues' astonishing documentation that "thousands of scientists publish a paper every five days."

But Must Publishing Be the Only Coin of the Realm?

Professors Nosek and Bar-Anan have a tacit answer for this question along with just about everything else associated with publishing. Namely (a variant of which has been suggested by others as well), that some "scientists who do not have the resources or interest in doing original research themselves can make substantial contributions to science by reviewing, rather than waiting to be asked to review" (p. 237).

In a sense all scientists are peer reviewers, if not as publication gatekeepers, at least for their own purposes every time they read an article relevant to their work. So why not officially create a profession given over to this activity, one accompanied by an official record of these activities for promotion and tenure purposes? Or, barring that, an increased and rewarded system of online reviews designed to discourage methodologically unsound, unoriginal, or absurd publications.

Alternately, there are presently multiple online sites upon which one's comments regarding the most egregious departures from good scientific practice can be shared with the profession as a whole and/or via one-on-one correspondences with the authors themselves. From a scientific perspective, if institutionalized as a legitimate academic discipline, the hopeful result of such activities would be to improve reproducibility one study and one investigator at a time.

There are, of course, many other options and professional models already proposed, such as Gary King's 1995 recommendation that scientists receive credit for the creation of datasets that facilitate the replication process. In models such as this scientists would be judged academically on their performance of duties designed to facilitate the scientific process itself, which could include a wide range of activities in addition to peer reviewing and the creation of databases. Already existing examples, such as research design experts and statisticians, are well established, but the list could be expanded to include checking preregistered protocols (including addendums thereto) against published or submitted final reports.

And, of course, given the number of publications being generated in every discipline there are abundant opportunities for spotting questionable research practices (QRPs) or errors in newly published studies. Perhaps not a particularly endearing professional role or profession, but letters to the offending journals' editors and/or postings on websites designed for the specific purpose of promulgating potential problems could be counted as

worthy, quantifiable professional activities. And, of course, the relatively new field of *meta-science* is presently open for candidates and undoubtedly has room for numerous subspecialties including the development of software to facilitate all of the just-mentioned activities. This is an activity which has already produced some quite impressive results, as described next.

Already Existing Statistical Tools to Facilitate These Roles and Purposes

Taking the search for potential errors and misconduct in published research as an example, there are a growing number of tools available to facilitate this purpose. In addition to ScholarOne, StatReviewer, and p-curve analysis already mentioned, other approaches for identifying *potential* statistical abnormalities exist—three of which are described clearly in a very informative article by Bergh, Sharp, and Li (2017) titled "Tests for Identifying 'Red Flags' in Empirical Findings: Demonstration and Recommendations for Authors, Reviewers, and Editors." A sampling of some other indirect (but creative approaches) for spotting statistical abnormalities included the following:

1. An R-program (statcheck) developed by Epskamp and Nuijten (2015) which allows extracted p-values to be recalculated to spot abnormalities, possible tampering, and errors based on reported descriptive statistics. Using this approach Nuijten, Hartgerink, van Assen, Epskamp, and Wicherts (2016) identified a disheartening number of incorrectly reported p-values in 16,695 published articles employing inferential statistics. As would be expected by now, substantively more false-positive errors than negative ones were found.
2. Simulations such as bootstrapping approaches (Goldfarb & King, 2016) for determining what would happen if a published research result were to be repeated numerous times, with each repetition being done with a new random draw of observations from the same underlying population.
3. A strategy (as described by Bergh, Sharp, Aguinis, & Li, 2017) offered by most statistical packages (e.g., Stata, IBM SPSS, SAS, and R) for checking the accuracy of statistical analyses involving descriptive and correlational results when raw data are not available. Using linear regression and structural equation modeling as examples, the authors

found that of those management studies for which sufficient data were available and hence could be reanalyzed, "nearly one of three reported hypotheses as statistically significant which were no longer so in retesting, and far more significant results were found to be non-significant in the reproductions than in the opposite direction" (p. 430).

4. The GRIM test (Brown & Heathers, 2016) which evaluates whether or not the summary statistics in a publication are mathematically possible based on sample size and number of items for whole (i.e., non-decimal) numbers, such as Likert scales.
5. Ulrich Schimmack's "test of insufficient variance" (2014) and "z-curve analysis" (Schimmack & Brunner, 2017) designed to detect QRPs and estimate replicability, respectively.

All of which could be a most propitious hobby or turn out to be an actual scientific discipline. By way of illustration, consider the following study emanating from an anesthesia researcher's hobby (or passion) for improving reproducibility in his chosen field.

Data Fabrication and Other Reasons for Non-Random Sampling in 5087 RCTs in Anesthetic and General Medical Journals

John B. Carlisle (2017)

Recently highlighted in a *Nature* article (Adam, 2019), Dr. Carlisle arises before dawn to let his cat out and begins entering published experimental anesthesia data into a spreadsheet that will eventually be analyzed for suspicious values—the presence of which he has the temerity to inform the editors of those journals in which the offending articles appear. A process, incidentally, that has resulted in the identification of both fraudulent investigators and numerous retractions.

The study being described here involved (as its title suggests) more than 5,000 anesthesia studies from six anesthesia and two general medical journals (the *Journal of the American Medical Association* and the *New England Journal of Medicine*). These latter two journals were most likely added because anesthesia research appears to be the medical analog to social psychology as far as suspicious activities are concerned. Hence

Dr. Carlisle may have wanted to ascertain if his profession did indeed constitute a medical outlier in this respect. (According to the *Nature* article, four anesthesia investigators [Yoshitaka Fujii, Yuhji Saitoh, Joachim Boldt, and Yoshihiro Sato] eventually had 392 articles retracted, which, according to Retraction Watch, dwarfs psychologist Diederik Stapel's 58 admitted data fabrications: https://retractionwatch.com/2015/12/08/diederik-stapel-now-has-58-retractions/.)

Carlisle's analysis involved 72,261 published arithmetic means of 29,789 variables in 5,087 trials. No significant difference occurred between anesthesia and general medicine with respect to their baseline value distributions, although the latter had a lower retraction rate than the former. And in agreement with just about all of the authors of this genre of research, Dr. Carlisle was quite explicit in stating that his results could not be interpreted as evidence of misconduct since they could also be functions of "unintentional error, correlation, stratified allocation and poor methodology."

He did implicitly suggest, however, that more investigators should join him in this enterprise since, "It is likely that this work will lead to the identification, correction and retraction of hitherto unretracted randomised, controlled trials" (p. 944).

And the *New England Journal of Medicine* obviously agrees since it has announced that it will be applying this technique to future submissions.

Relatedly, Bolland, Avenell, and Gamble (2016) applied a variant of Dr. Carlisle's approach via a meta-analysis of 33 problematic trials investigating elderly falls (i.e., problematic with respect to baseline data "involving [uncharacteristically] large numbers of older patients with substantial comorbidity, recruited over very short periods"). The results of that analysis being that

> [o]utcomes were remarkably positive, with very low mortality and study withdrawals despite substantial comorbidity. There were very large reductions in hip fracture incidence, regardless of intervention (relative risk 0.22, 95% confidence interval 0.15–0.31, $p < 0.0001$ . . . that greatly exceed those reported in meta-analyses of other trials. There were multiple examples of inconsistencies between and within trials, errors in reported data, misleading text, duplicated data and text, and uncertainties about ethical oversight. (p. 1)

So Is It Time for Publishers' to Step Up?

It is past time, regardless of how the journal system evolves over the next few decades. Hopefully the *New England Journal of Medicine*'s tentative step in following John Carlisle's lead is only a precursor to more hands-on actions by journals to ensure the integrity of what they publish. Perhaps the reproducibility crisis's expanding profile will facilitate such actions, supplemented by continuing efforts by scientists such as Dr. Carlisle and the determined efforts of the myriad contributors to this scientific initiative. For it is important to remember that scientific publishing is not solely in the hands of the publishing industry CEOs and CFOs or the editors-in-chief. Rather it is a symbiotic process involving multiple other actors, the most important of which are scientists themselves.

Methodological recommendations in these regards have been tendered by a number of the reproducibility advocates mentioned in recent chapters so there is no need to restate them here. However, given the descriptions of statistical/empirical aids for ensuring the validity of empirical results just mentioned, let's revisit the aforementioned Bergh et al.'s "Red Flags" article that very succinctly reminds us of the multiple roles and responsibilities for ensuring reproducibility from a statistical perspective in the initial phase of the publishing process:

First author's responsibilities:

1. Include such values as coefficient estimates, standard errors, p-values in decimals, and a correlation matrix that includes means, standard deviations, correlations [including those between covariate and outcome], and sample sizes.
2. "Describe all data-related decisions such as transformed variables and how missing values and outliers were handled."
3. "Attest to the accuracy of the data and that the reporting of analytical findings and conclusions."

Next, the editor's responsibilities:

1. Ensure that all of the previous disclosure requirements are satisfied.
2. Require "authors to attest that their findings are based on the reported data and analytical findings; indicate that findings will be confirmed through retesting if article receives a conditional acceptance."

3. "Amend manuscript evaluation form sent to reviewers to include a check of the expanded data disclosure reporting requirements and for consistency between disclosure, analysis, hypotheses, and conclusions."
4. "Retest findings using Tests 1 [ensuring the accuracy of p-values] and 3 [verifying study results based on matrices of descriptive statistic] after a conditional acceptance is awarded and before a final acceptance is reached." (Obviously, the editors themselves won't do this but will task an employee or contractor to do so.)

And, of course, the peer reviewers' responsibilities:

1. "Confirm that data reporting is complete and meets expanded disclosure requirements (permitting the tests described above)."
2. "Assess relationships between the data, findings, and interpretation of hypotheses to ensure consistency." (All dicta are courtesy of table 5, p. 122—exact quotations are so noted, others are paraphrased.)

Plus a few from Philip Bourne and Alon Korngreen (2006):

1. Do not accept a review assignment unless you can accomplish the task in the requested timeframe—learn to say no.
2. Avoid conflict of interest [e.g., cronyism].
3. As a reviewer you are part of the authoring process [which means you should strive to make whatever you review a better paper].
4. Spend your precious time on papers worthy of a good review.
5. Write clearly, succinctly, and in a neutral tone, but be decisive.
6. Make use of the "comments to editors" [i.e., comments designed to help the editor but not to be shared with the author] (pp. 0973–0974).

A Final Publishing Vision that Should Indirectly Improve Reproducibility

Reading the following brief article when it was first published would probably have been viewed as absurd by most social scientists dabbling in computational research (and completely irrelevant for those not involved therein). Today, however, in the context of the reproducibility crisis, it resonates as

downright prescient for improving all genres of scientists' day-to-day empirical practice.

What Do I Want From the Publisher of the Future?

Philip Bourne (2010)

Written by the editor of *PLoS Computational Biology*, in some ways this editorial picks up where the just discussed article by Nosek and Bar-Anan leaves off even though it was published 2 years earlier. The paper would be greatly improved by a few detailed examples, but it isn't fair to criticize a brief editorial targeted at computational scientists on this basis since undoubtedly almost everyone in that discipline could supply several such examples from their own experience.

So, in support of many of Drs. Nosek and Bar-Anan's suggested changes, consider the following rhetorical question posed by Dr. Bourne:

> After all our efforts at producing a paper, very few of us have asked the question, is journal *X* presenting my work in a way that maximizes the understanding of what has been done, *providing the means to ensure maximum reproducibility* [emphasis added] of what has been done, and maximizing the outreach of my work? (p. 1)

Dr. Bourne's answer to this question is a resounding "no," consonant with Nosek and Bar-Anan's (2012) observation:

> Authors are so happy to have their submission accepted that they blissfully sign a copyright transfer form sent by the publishers. Then publishers recoup their investment by *closing* [emphasis added] access to the articles and then selling journal subscriptions to the scientists and their institutions (individual articles can be purchased for $5 to $50 depending on the journal). [Actually, in my experience it is a rare article that can be obtained as cheaply as $5.] In other words, the funding public, universities, and scientists who produced and pay for the research give ownership of the results to publishers. Then, those with money left over buy the results back from the publishers; the rest are in the dark. (p. 228)

Now back to Professor Bourne, who suggests that, in addition to what is normally included in a scientific article (i.e., the stored data and its documentation—recalling that his specialty is computational research where usable shared data is usually a publication *requirement*), journals (or third parties associated therewith) should also publish what he calls the relevant laboratory's *workflow*—which constitutes a detailed record of everything pertinent to the creation of the published or deposited article.

Naturally, such a workflow might be quite different for experimental work involving humans and animals. However, every laboratory and every study should keep a detailed workflow comprising every aspect of a study from the initial idea to the details of the literature review to descriptions of any meetings or correspondence with investigators or assistants to the writing of the regulatory and/or funding proposals. Pilot studies should also be recorded in detail, including accrued data as well as for those aborted along the way. And, of course, every aspect of the actual study would be available therein, including

1. The final registered protocol;
2. The approved institutional review board (IRB) or institutional animal care and use committee (IACUC) proposal (including correspondences regarding it and any amendments submitted thereto);
3. Any procedural glitches and/or minor "tweaking" of study conditions, outcome variables, and analytic decisions; and
4. All correspondences with journal editors and peer reviewers.

While Dr. Bourne's vision of a computational workflow would probably not be quite this comprehensive, he provides some universal advantages of the process based on correctable inefficiencies experienced in his own lab.

1. The intellectual memory of my laboratory is in my e-mail folders, themselves not perfectly organized. This creates a hub-and-spoke environment where lab members and collaborators have to too often go through me to connect to each other.
2. Much of our outreach is in the form of presentations made to each other and at national and international forums. We do not have a good central repository for this material; such a repository could enable us to have a better understanding of what other researchers are doing.

3. While we endeavor to make all our software open source, there are always useful bits of code that languish and disappear when the author leaves the laboratory.
4. Important data get lost as students and postdoctoral fellows leave the laboratory. (p. 2)

Now certainly most workflows will probably never be archived by a publisher. However, there is no reason why individuals cannot and should not keep better, even compulsively detailed, *personal* records of their research for at least two very rather obvious reasons.

First, comprehensive workflows will greatly facilitate the replication process because even if a request for information from an independent replicator of one's study is stressful, it is advantageous for the original investigator that the replication be performed with maximum fidelity. (And, has been mentioned previously, a direct replication is seldom possible based on nothing more than the information typically available in the methods section of a typical research article.)

Second, when investigators are young and have conducted only a few studies, their memories may be sufficient to reconstruct the exact procedures employed in their studies. However, with time and its many other competitors for cerebral storage, earlier memories begin to deteriorate—a phenomenon for which I could serve as an expert witness.

A parable: As a bizarre hypothetical example, let's return to our hypothetical graduate student of long and many pages ago and his obviously fictitious series of 20+ incremental experiments (all negative) designed to develop a prose study skill capable of producing superior learning to reading and rereading a passage when *time on task was held constant*. As we know, to his everlasting shame, he never attempted to publish this rather Quixotic effort since none of the experiments reached statistical significance (recalling that publication bias of this sort is a very old phenomenon, as the graduate student would be by now had he ever existed).

Suppose further that within a couple of decades he had realized the folly of his ways because, by then, his results appeared to have significant implications for a time-on-task theory he was working on. A theory, in fact, that not coincidentally explained his earlier failures to produce an efficacious method of study and could be validated by a relatively simple series of experiments.

However, there was no possibility that he could recall sufficient details concerning even a few of the studies, much less all of them (or even how many he had conducted). And, of course, practically no one in those days kept detailed paper-based records of procedures, detailed protocols, or data for decades, especially following the institutional moves that often accompany such time intervals—and especially not for unpublished studies. (Hence, even the file drawer constitutes an insufficient metaphor for this problem.)

Today, however, after the Digital Revolution, there is really no excuse for such behavior. There is also no excuse for not sharing all of the information contributing to and surrounding a scientific finding, published or unpublished. And that just happens to constitute the subject of Chapter 10.

References

Adam, D. (2019). The data detective. *Nature, 571*, 462–464.

Allison, D. B., Brown, A. W., George, B. J., & Kaiser, K. A. (2016). A tragedy of errors: Mistakes in peer reviewed papers are easy to find but hard to fix. *Nature, 530*, 27–29.

Baldwin, R. (2018) Retracted Article: Who are they to judge? Overcoming anthropometry through fat bodybuilding, *Fat Studies, 7*, i–xiii.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization, 15*, 423–436.

Bergh, D. D., Sharp, B. M., & Li, M. (2017). Tests for identifying "Red Flags" in empirical findings: Demonstration and recommendations for authors, reviewers, and editors. *Academy of Management Learning and Education*, *16*, 110–124.

Björk, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research, 14*(1), paper 391.

Bohannon, J. (2013). Who's afraid of peer review? *Science, 342*, 60–65.

Bolland, M. J., Avenell, A., & Gamble, G. D. (2016). Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. *Neurology, 87*, 1–12.

Bourne, P. E. (2010). What do I want from the publisher of the future? *PLoS Computational Biology*, *6*, e1000787.

Bourne P. E., & Korngreen, A. (2006). Ten simple rules for reviewers. *PLoS Computational Biology, 2*, e110.

Brown, N. J., & Heathers, J. A. (2016). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, https://peerj.com/preprints/2064.pdf.

Butler, D. (2013). The dark side of publishing. The explosion in open-access publishing has enabled the rise of questionable operators. *Nature, 435*, 433–435.

Carlisle, J. B. (2017). Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*, *72*, 944–952.

Epskamp, S., & Nuijten, M. B. (2015). Statcheck: Extract statistics from articles and recompute p values. R package version 1.0.1. http://CRAN.R-project.org/package=statcheck

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Science, A, 109*, 17028–17033.

Ferguson, C., Marcus, A., & Oransky, I. (2014). Publishing: The peer review scam. *Nature, 515*, 480–482.

Goldfarb, B. D., & King, A. A. (2016). Scientific apophenia in strategic management research: Significance tests and mistaken inference. *Strategic Management Journal, 37*, 167–176.

Heaven, D. (2018). AI peer reviewers unleashed to ease publishing grind. *Nature, 563*, 609–610.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*, e124.

Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2018). Thousands of scientists publish a paper every five days, papers and trying to understand what the authors have done. *Nature, 561*, 167–169.

King, G. (1995). Replication, replication. *PS: Political Science and Politics, 28*, 444–452.

Mallapaty, S. (2020). Popular preprint sites face closure because of money troubles. *Nature, 578*, 349.

Martinson, B. C. (2017). Give researchers a lifetime word limit. *Nature, 550*, 202.

National Science Board. (2018). *Science and engineering indicators 2018*. NSB-2018-1. Alexandria, VA: National Science Foundation. (www.nsf.gov/statistics/indicators/)

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry, 23*, 291–293.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia I: Opening scientific communication. *Psychological Inquiry, 23*, 217–243.

Nuijten, M., Hartgerink, C. J., van Assen, M. L. M., Epskamp, S., & Wicherts, J. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*, 1205–1226.

Schimmack, U. (2014). The test of insufficient variance (TIVA): A new tool for the detection of questionable research practices. https://replicationindex.wordpress.com/2014/12/30/the-test-ofinsufficientvariance-tiva-a-new-tool-for-the-detection-ofquestionableresearch-practices/

Schimmack, U., & Brunner, J. (2017). Z-curve: A method for estimating replicability based on test statistics in original studies. https://replicationindex.files.wordpress.com/2017/11/z-curve-submission-draft.pdf

Sokal, A. D. (1996). Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity. *Social Text*, *46–47*, 217–252,

Tsang, E. W., & Frey, B. S. (2007). The as-is journal review process: Let authors own their ideas. *Academy of Management Learning and Education*, *6*, 128–136.

Wilson, H. (2018). Retracted Article: Human reactions to rape culture and queer performativity at urban dog parks in Portland, Oregon. *Gender, Place & Culture, 27*(2), 1–20.

10

Preregistration, Data Sharing, and Other Salutary Behaviors

The purpose of this chapter is to discuss some of the most powerful, effective, and painless strategies for lowering the prevalence of irreproducible scientific findings in a bit more depth. So hopefully I will be excused for a bit of redundancy here given the crucial role preregistration and data sharing play in reducing the opportunities for poorly trained or inadequately acculturated or even unscrupulous scientists to engage in research driven by questionable research practices (QRPs).

For while it is unfortunate that these individuals have been denied (or have chosen to ignore) educational or principled mentoring opportunities, it is absolutely necessary that the continued production of false-positive results in the scientific literature be arrested. And it is highly unlikely that the continuing promulgation of explicit edicts regarding the behaviors in question will be capable of trumping these individuals' perceived (or real) need to continue to publish such findings.

So, in way of review, what are needed are universal requirements for

1. The protocols of all empirical research to be preregistered accompanied by amendments detailing any procedural changes occurring during the conduct thereof, and
2. All data accruing from these protocols to be deposited in an accessible, discipline-approved registry (i.e., not a personal website) and available to all scientists, along with all the code required to permit it to be quickly run as a check against the published descriptive and inferential results.

However, as will be illustrated shortly, this, too, will be insufficient if

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0011

1. The preregistered protocols are not compared point-by-point to their published counterparts (preferably involving a standardized methodological and statistical checklist), and
2. The registered data are not reanalyzed (using the available code) as a descriptive check on the demographics and as an inferential check on the primary p-values reported.

But who is going to do all of this? The easy answer is the multi-billion dollar publishing industry, but even though its flagship, the *New England Journal of Medicine*, has joined Dr. Carlisle in checking distributions of the baseline values of submitted randomized controlled trials (RCTs), this does not mean that other journals will embrace this strategy (or the considerably more effective and time-consuming suggestion just tendered of reanalyzing the actual data).

However, since this industry is built on the free labor of scientists, why not create a profession (as alluded to in Chapter 9) devoted to checking preregistered protocols against the submitted research reports and actually rerunning the key analyses based on the registered data as part of the peer review process? These efforts could be acknowledged via (a) a footnote in the published articles; (b) journal backmatter volume lists, in which peer reviewers and authors are often listed; (c) or even rewarded as a new form of "confirming" authorship. (And, naturally, these behaviors would be acknowledged and rewarded by the participants' institutions.)

And why is all of this necessary since replication is the reproducibility gold standard? The easy answer is that the number of studies presently being published precludes replicating even a small fraction thereof given the resources required. (Not to mention the fact that some studies cannot be replicated for various reasons.)

So, if nothing else, with a study that passes these initial screenings involving the original protocol and the registered data (and assuming it also passes the peer review process), considerable more confidence can be had in its ultimate reproducibility. This evidence, coupled with the perceived importance of a study, will also help inform whether a replication is called for or not. And if it is, the registered information will make the replication process considerably easier—for not only is it the gold standard for reproducibility, its increased practice provides a powerful disincentive for conducting substandard research.

But if a study protocol is preregistered why is it necessary to check it against the final published product? Partly because of scientists' unique status in society and the conditions under which they work. Unlike many professions, while scientists typically work in teams, their individual behaviors are not supervised. If fact, most senior scientists do not actually perform any of the procedural behaviors involved in their experiments but instead rely on giving instructions to research assistants or postdocs who themselves are rarely supervised on a day-to-day basis—at least following a training run or two.

Of course scientists, like other professionals such as surgeons, are expected to follow explicit, evidence-based guidelines. But while surgeons are rewarded for positive results, their operating procedures can also be investigated based on an egregious and expected negative result (or a confluence thereof via malpractice suits or facing the families of patients who have died under their watch)—*unless* they have rigorously followed evidence-based, standard operating (not a pun) procedures. And while scientists also have standard operating procedures, as documented in guidelines such as the CONSORT and ARRIVE statements, their carrots and sticks are quite different. The "sticks" have historically come in the form of little more than an occasional carrot reduction, with the exception of the commission of outright fraud. (And some countries even permit investigators who have committed egregious examples of fraud to continue to practice and publish.) The carrots come in the form of tenure, direct deposit increases, and the esteem (or jealousy) of their colleagues, which in turn requires numerous publications, significant external funding, and, of course statistically significant research results.

All of which *may* have resulted in the social and life sciences waking up to the reproducibility crisis after a prolonged nap during which

1. The scientific version of standard operating procedures involved avoiding some of the QRPs listed in Chapter 3 while ignoring many of the others; hence,
2. Scientific training had typically underplayed the problematic nature of these practices; while
3. There has been increased pressure from the publishing industry to keep research reports brief in order to include as many articles as possible in each issue; with
4. Little or no cultural imperative for change for any of the preceding via the "if it isn't broken don't fix it" paradigm.

But, of course, "it" is broken, and the traditional emphasis on intensive practice as the primary mechanism for teaching the skills necessary for the continuity of the scientific professions is not sufficient. At least, not if this intensive practice involves the continuance of traditional research behaviors. Thus the most important message emanating from the reproducibility movement is that many (not all) of the traditional ways in which science has been conducted must change and that the most important mechanisms to drive this change involve *required* transparency in the conduct of the research practices which *precede* the collection of a study's first data point. (Or, for investigations involving existing databases, transparent steps that *precede* inferential analyses.)

But barring constantly monitored surveillance equipment in the laboratory, which isn't a bad idea in some cases (as recommended by Timothy Clark [2017]), what could accomplish such a task? Hopefully the requisite behaviors are obvious by now, but let's examine these behaviors' absolute essentials, their requisite policies, and their current levels of implementation in a bit more detail.

Preregistration of Study Protocols

The primary purpose of the preregistration of research protocols is to prevent the real and simulated methodological abuses detailed in previous chapters. The Simmons, Nelson, and Simonsohn "Hot Potatoes" study (2011), for example, would have most likely been accepted for publication in the past as initially written sans any disclaimers. However, even in the "good old days," had the study's hypotheses, experimental conditions, primary outcomes, sample size justification, inclusion/exclusion criteria, and analytic approaches (such as the handling of missing data and the use of covariates and blocking variables) been preregistered on an accessible website prior to running the first participant and that preregistered document compared to the submitted publication by either a peer reviewer or an employee of the publisher, then the submitted manuscript would have surely been summarily rejected.

However, even if no officially registered protocol–manuscript comparison had taken place, the process itself would serve a number of useful purposes.

1. The very *threat* of third-party preregistration evaluations broadcasting any discrepancies on social media might be a significant deterrent in and of itself;
2. The preregistration document could serve as a reference to conscientious investigators in the preparation of their final manuscripts by reminding them of exactly what they had originally proposed and what actually transpired in the conduct of the study—especially when there is a significant delay between study initiation and manuscript preparation;
3. Since changes and amendments to the document following study commencement are often necessary, the preregistration process and the highly recommended laboratory workflow can operate together synergistically to keep track of progress and facilitate memories.

In the long run, then, the very existence of a thorough preregistration document accompanied by amendments should serve to encourage investigators to transparently list major changes in the published document occurring between the originally proposed and completed study. This, in turn, will facilitate publication by stealing peer reviewers' and critics' thunder and, in some instances, turn a potential QRP into a non sequitur. And finally, the very threat of future methodological criticisms might make some investigators more careful in the conduct of their studies and their subsequent writing style.

The Initiation of the Preregistration Movement

The most substantive move toward requiring preregistration as a condition of publication came from journal editors themselves in 2004, after several years of sporadic lobbying in the form of articles by methodologists and, in some cases, professional associations. In that year the International Committee of Medical Journal Editors (ICMJE), which is comprised of the most prestigious journals in the field (including the *New England Journal of Medicine* and the *Journal of the American Medical Association*) announced that, as of July 1, 2005, preregistration of all clinical *trials* would become a prerequisite for publication in all of the organizations' affiliated journals (De Angelis, Drazen, Frizelle, et al., 2004).

Interestingly, the primary rationale for instituting this policy appeared to be to vitiate publication bias, which has a somewhat different genesis (and an even more important implication) in medicine as opposed to the social sciences. Namely, it was undertaken to prevent research sponsors such as the pharmaceutical industry (second only to the federal government as a source of medical research funding) from concealing the presence of selected (presumably negative) trials that could potentially "influence the thinking of patients, clinicians, other researchers, and experts who write practice guidelines or decide on insurance-coverage policy" (De Angelis et al., 2004, p. 1250).

The early requirements for registries were actually rather modest and could be of the investigators' choosing as long as they were (a) free of charge and accessible to the public, (b) open to all prospective registrants, and (c) managed by a not-for-profit organization. The requirements for the preregistrations of the clinical trials themselves, while far from onerous, suggested an acute awareness of many of the QRPs listed in Chapters 3 and 4.

> There must be a mechanism to ensure the validity of the registration data, and the registry should be electronically searchable. An acceptable registry must include at minimum the following information: a unique identifying number, a statement of the intervention (or interventions) and comparison (or comparisons) studied, a statement of the study hypothesis, definitions of the primary and secondary outcome measures, eligibility criteria, key trial dates (registration date, anticipated or actual start date, anticipated or actual date of last follow-up, planned or actual date of closure to data entry, and date trial data considered complete), target number of subjects, funding source, and contact information for the principal investigator. (p. 1251)

The social sciences eventually followed suit. The Open Science Project undoubtedly provided the greatest impetus and, as of this writing, has more than 8,000 registered studies from a wide variety of disciplines. There are also an unknown but large number of US and international research registries in a bewildering number of disciplines—*clinicaltrials.com* being the largest with more than 300,000 registered trials representing more than 200 countries. (The World Health Organization's International Clinical Trials Registry Platform is also huge, listing approximately 200 participating countries.)

Epistemologically, the most important functions served by the preregistration of trials may be the potential of the process to operationally differentiate between hypotheses (a) generated a posteriori based on a completed a study's data and (b) those generated a priori and constituting the actual rationale for the study in the first place. Obviously a great deal more confidence can be had in the latter as opposed to the former—perhaps most effectively illustrated metaphorically by James Mills (1993) in his classic article "Data Torturing" ("If the fishing expedition catches a boot, the fishermen should throw it back, not claim that they were fishing for boots" [p. 1198]) and Andrew Gelman and Eric Loken's (2014) previously discussed seminal article (in a "garden of forking paths, whatever route you take seems predetermined" [p. 464]).

This genre of ex post facto behavior is surely one of the most frequently engaged in of QRPs (our number 15 in Chapter 3), one of the most insidious, and more often than not accompanied by several related undesirable practices. It is also a leading cause of irreproducibility and publication bias. And, like most QRBs, its end results (artifactually produced and incorrect p-values < 0.05) are reinforced by publication and peer review policies.

However, while preregistration is an important preventive measure for ex post factor hypothesizing, the practice has a long history in the social sciences. So let's take a quick trip back in time and review a small piece of the history of this particular QRP—and in the process perhaps help explain why it is so difficult to eradicate.

HARKing: Hypothesizing After the Results Are Known

Norbert Kerr (1991)

This is a classic and even-handed discussion of what Brian Nosek and colleagues (2018) would later refer to as "postdiction" and to which Professor Kerr bestowed the acronym "HARKing" before concluding that its disadvantages far outweighed its few advantages and hence was completely contraindicated unless an unpredicted-unexpected finding was clearly labeled as such. However, an almost equally interesting aspect of the article is its illustration of how widely accepted the practice apparently was in 1991. Even such a methodological luminary as Daryl Bem is quoted as giving the following advice to students and new researchers in a book chapter entitled "Writing the Empirical Journal Article." (The 1987

book itself was designed to mentor the soon to be practicing generation of researchers and was aptly titled *The Compleat Academic: A Practical Guide for the Beginning Social Scientist*.)

Sample Bem quotes regarding HARKing include the following:

> There are two possible articles you can write: (1) the article you planned to write when you designed your study or (2) *the article that makes the most sense now that you have seen the results* [emphasis added]. They are rarely the same, and the correct answer is (2) . . . the best journal articles are informed by the actual empirical findings from the opening sentence. (pp. 171–172)

Or,

> The data may be strong enough to justify recentering your article around the new findings and subordinating or even ignoring your original hypotheses. . . . If your results suggest a compelling framework for their presentation, adopt it and make the most instructive findings your centerpiece. (p. 173)

Now it's easy to bash Daryl Bem today, but this sort of attitude, approach, or orientation toward ensuring publication through either writing a research report or conducting the research seemed to be generally acceptable a quarter of a century ago. This is better illustrated via an unpublished survey Professor Kerr conducted in 1991 (with S. E. Harris), in which 156 behavioral scientists were asked to estimate the frequency that they suspected HARKing (somewhat broadly defined) was practiced in their discipline. A majority reported the belief that this set of behaviors was actually practiced more frequently than the classic approach to hypothesis testing. Finally, Professor Kerr advanced a number of untoward side effects of HARking including (in his words)

1. Translating Type I errors into hard-to-eradicate theory;
2. Propounding theories that cannot (pending replication) pass Popper's disconfirmability test;
3. Disguising post hoc explanations as a priori explanations (when the former tend also be more ad hoc, and consequently, less useful);
4. Not communicating valuable information about what did not work;
5. Taking unjustified statistical license;
6. Presenting an inaccurate model of science to students;

7. Encouraging "fudging" in other grey areas;
8. Making us less receptive to serendipitous findings;
9. Encouraging adoption of narrow, context-bound new theory;
10. Encouraging retention of too-broad, disconfirmable old theory;
11. Inhibiting identification of plausible alternative hypotheses;

And last but definitely not least:

12. Implicitly violating basic ethical principles. (p. 211).

This absolute gem of an article is reminiscent of Anthony Greenwald's (1975) previously discussed classic. One almost feels as though the social sciences in general (and the reproducibility crisis in specific) are in some sort of bizarre time loop where everything is repeated every few decades with little more than a change of terminology and an escalating number of publishing opportunities.

The Current State of the Preregistration Movement

While Anthony Greenwald and Norbert Kerr are hard acts to follow, someone must carry on the tradition since few practicing researchers read or heed what anyone has written or advocated decades in the past. And certainly Brian Nosek (and his numerous collaborators) is eminently qualified to assume that mantle as witnessed by the establishment of the Open Science Collaboration, which advocates preregistration and provides a multidisciplinary registry of its own, and a number of instructional articles on the topic (e.g., Nosek & Lakens, 2014) as well as the following aptly titled article.

The Preregistration Revolution

Brian Nosek, Charles Ebersole, Alexander DeHaven, and David Mellor (2018)

As just about everyone interested in scientific reproducibility agrees, one of the most effective ways of discouraging scientists from disguising

postdiction as prediction is to require them to publicly register their primary hypotheses and approaches prior to data collection. However, Professor Nosek and his co-authors are quick to emphasize, as did Kerr before them, that both prediction and postdiction are integral parts of science and that "Preregistration does not favor prediction over postdiction; its purpose is to make clear which is which" (p. 2602).

The authors make a clear distinction between generating a hypothesis and then testing it with data versus "hypothesizing after the results are known" (with a well-deserved nod to Professor Kerr). And while the untoward effects of this form of illogical reasoning and faulty empirical practice have already been illustrated in several ways, this particular article provides a more succinct description of the underlying logic behind this most serious and pervasive of artifacts.

> It is an example of circular reasoning––generating a hypothesis based on observing data, and then evaluating the validity of the hypothesis based on the same data. (p. 2600)

The authors then go on to list nine challenges that can be associated with the preregistration process. The first seven will be discussed here because they are seldom (if ever) mentioned in the context of preregistration, although the description here does not do them justice and they are better read in their entirety in the original article.

1. *Changes to procedure during the conduct of the study*. Probably the most common example of this occurs in clinical trials when recruitment turns out to be more difficult than anticipated. To paraphrase a shock trauma investigator I once worked with, "The best prevention for spinal cord injury is to conduct a clinical trial requiring the recruitment of spinal cord injured patients. Diving and motor cycle accidents will inevitably almost completely disappear as soon as study recruitment begins." Which, of course, sometimes unavoidably results in readjusting the study's originally proposed sample size. But this is only one of a multitude of other unanticipated glitches or necessary changes to a protocol that can occur after the study begins. As the authors suggest, transparently reporting what these changes were and the reason that they were necessitated will go a long way toward salvaging a study and making it useful, *unless*

said changes were made after looking at the data (with the exception of the next "challenge").

2. *Discovery of assumption violations during analysis*. This one is best avoided by pilot work, but distribution violations may occur with a larger and/or slightly different sample than was used in the pilot study process. The authors suggest that a "decision tree" approach be specified in the preregistration regarding what analytic steps will be taken if the data do not fit the preregistered analytic plan (e.g., non-normality or missing values on a prespecified covariate). However, some unanticipated problems (e.g., a ceiling effect or basement effect with regard to the outcome variable) can be fatal, so it should be noted that the options presented deal only with the violation of statistical assumptions. By the same token, all experienced investigators have a pretty thorough knowledge of what *could* possibly occur during the course of a study in their fields, hence Lin and Green's (2016) suggestion that common genres of research adopt SOPs which can be copied and pasted into a preregistration document to cover possible discrepancies between published and prespecified analysis plans.
3. *Analyses based upon preexisting data* and (4) *longitudinal studies and large, multivariate databases*. These two research genres involve uses of preregistration that are seldom considered. For example, the utility of blinding is well-established in experimental research but it can also apply to longitudinal databases in the form of generating and registering hypotheses prior to data analysis. When this isn't practical perhaps a reasonable fallback position would be to include those relationships which the investigators have already discovered and reported in a preregistration document while transparently reporting them as such in the published analysis, accompanied by an alpha adjustment of 0.005. (Which isn't as onerous for large databases as it is for experiments.) And, of course, non-hypothesized relationships found to be of interest should also be similarly reported. (These latter suggestions shouldn't be attributed to the Nosek team since they are my opinions.)
5. *Running many experiments at the same time*. Here, the authors described a situation in which a "laboratory acquires data quickly, sometimes running multiple experiments per week. The notion of pre-registering every experiment seems highly burdensome for

their efficient workflow" (p. 2603). Some might find this scenario a bit troublesome since it is highly unlikely that said laboratories publish all of these "multiple experiments per week" hence the trashing of the non-significant ones might constitute a QRP in and of itself. In their defense, the authors suggest that this is normally done "in the context of a methodological paradigm in which each experiment varies some key aspects of a common procedure" (p. 2603). So, in this case, the authors describe how a preregistration can be written for such a program of research as a whole, and any promising findings can then be replicated. However, for experiments conducted in this manner which do not simply vary "some key aspects of a common procedure," one wonders what happens to all of the "negative" findings. Are they never published? At the very least, they should at least be mentioned in the published article and recorded in the laboratory's official workflow.

6. *Conducting a program of research*. This one is a bit like the previous challenge but seems to involve a series of separate full-blown experiments, each of which is preregistered and one eventually turns out to be statistically significant at the 0.05 level. The authors make the important point that such an investigator (reminiscent of our hypothetical graduate student's failed program of research) should report the number of failures preceding his or her success since the latter is more likely to be a chance finding than a stand-alone study. Few investigators would either consider doing this or adjusting the positive p-value based on the number of previous failures. But they probably should and, perhaps in the future, will.
7. *Conducting "discovery" research with no actual hypotheses*. In this scenario, similar in some ways to challenges (3) and (4), researchers freely admit that their research is exploratory and hence may see no need to preregister said studies. The authors conclude that this could be quite reasonable, but it is a process fraught with dangers, one of which is that it is quite possible for scientists to fool themselves and truly believe that an exciting new finding was indeed suspected all along (i.e., "the garden of forking paths"). Preregistration guards against this possibility (or others' suspicions thereof) and possesses a number of other advantages as well—serving as an imperfect genre of workflow for investigators who do not routinely keep one.

The authors conclude their truly splendid article by suggesting that the preregistration movement appears to be accelerating, as illustrated by the numbers of existing research registries across an impressive number of disciplines and organizations. They also list resources designed to facilitate the process, including online courses and publishing incentives, while warning that the movement still has a long way to go before it is a universal scientific norm.

So why not use this article to draw a red line for reproducibility? I have suggested tolerance for investigators such as Carney and Cuddy because they were simply doing what their colleagues had been doing for decades. Some have even partially excused Daryl Bem by saying that his work met minimum methodological standards for its times, but let's not go that far since there is really no excuse for pathological science.

But surely, at some point, enough becomes enough. Perhaps a year after the publication date of the preceding article (2018) could constitute such a red line. A zero tolerance point, if you will, one beyond which it is no longer necessary to be kindly or politically correct or civil in copious correspondences to offending editors and investigators, nor on posts on social media regarding studies published beyond this point in time that ignore (or its authors are ignorant of) the precepts that have been laid down from Nosek et al.'s 2018 declaration of a "Preregistration Revolution." And this is not to mention the myriad publications that surfaced around 2011–2012 or Anthony Greenwald's 1975 classic.

But isn't this a bit Draconian, given the progress we're making (or is something missing)? There is definitely something missing, even from medicine's inspired edicts listing the registration of clinical trials as a publication *requirement* in its highest impact journals after 2004—and even after this became a legal requirement for some types of clinical trials in 2007, with the law being expanded in 2017.

The first problem, as documented by a number of studies, involves the lack of compliance with preregistration edict. The second reflects the far too common mismatch between the preregistered protocol and what was actually published.

Mathieu, Boutron, Moher, and colleagues (2009) illustrated the necessity for ameliorating both of these problems in a study designed to compare the key elements present in preregistrations with their published counterpart.

Locating 323 cardiology, rheumatology, and gastroenterology trials published in 10 medical journals in 2008, these investigators found that only 147 (46%) had been adequately registered *before the end of the trial* despite the International Committee of Medical Journal Editors 2004 edict. And almost equally disheartening, 46 (31%) of these 147 compliant articles showed a discrepancy in the primary outcome specifications between the registered and published outcomes. And definitely most problematically, 19 (83%) of the 23 studies for which the direction of the discrepancies could be assessed were associated with statistically significant results. These published versus preregistration discrepancies were distributed as follows:

1. The introduction of a new primary outcome in the article (N = 22 studies);
2. The registered primary outcome was not even mentioned in the article (N = 15, which borders on the unbelievable);
3. The primary outcome morphing into a secondary one (i.e., either a secondary outcome or an unregistered variable becoming the primary outcome in the published article; N = 8);
4. A secondary outcome morphing into a primary one (N = 6); and/or
5. A discrepancy in the timing of the primary outcome between the two sources (N = 4).

(Note that some of the 46 articles had more than one of these QRPs.)

In an interesting coincidence, in the same publication year as this study, Ewart, Lausen, and Millian (2009) performed an analysis of 110 clinical trials and found the same percentage (31%) of primary outcomes changed from preregistration to published article, while fully 70% of the registered secondary outcomes had also been changed.

And, 2 years later, Huić, Marušić, and Marušić (2011) conducted a similar study comparing a set of 152 published RCTs registered in ClinicalTrials.gov with respect to both completeness and substantive discrepancies between the registry entries and the published reports. As would be expected by now, missing fields were found in the preregistrations themselves as well as substantive changes in the primary outcome (17%). Progress from 31% perhaps?

Now granted this is a lot of number parsing, but for those whose eyes have glazed over, suffice it to say that while the ICMJE initiative was a seminally exemplary, long overdue, and bordering on revolutionary policy for a tradition-bound discipline such as medicine and undeniably better than

nothing, it was, however, quite disappointing in the compliance it elicited several years after initiation.

Naturally all of the investigators just discussed had suggestions for improving the preregistration process, most of which should sound familiar by now. From Mathieu and colleagues (2009):

> First, the sponsor and principal investigator should ensure that the trial details are registered *before* [emphasis added] enrolling participants.
>
> Second. the comprehensiveness of the registration should be routinely checked by editors and readers, especially regarding the adequate reporting of important items such as the primary outcome.
>
> Third, editors and peer reviewers should systematically check the consistency between the registered protocol and the submitted manuscript to identify any discrepancies and, if necessary, require explanations from the authors, and
>
> Finally, the goal of trial registration could [*should*] be to make available and visible information about the existence and design of any trial and give full access to all trial protocols and the main trial results. (p. 984)

The conclusions for Huić et al., on the other hand, were a combination of good and bad news.

> ICMJE journals published RCTs with proper registration [the good news] but the registration data were often not adequate, underwent substantial changes in the registry over time and differed in registered and published data [the *very* bad news]. Editors need to establish quality control procedures in the journals so that they continue to contribute to the increased transparency of clinical trials. (p. 1)

In summary: While the movement toward preregistration is an inspired and profound boost to scientific reproducibility, it should in no way be voluntary. It should be required in some form or another for every published study—not just in medicine but in all disciplines aspiring to be taken seriously. From a journal's perspective, this process might entail the following steps:

1. Since manuscripts are now submitted online, the first item on the submission form should include a direct link to the *dated* preregistration

document, and the submission process should be summarily terminated if that field is missing. Any major deviations from the preregistration document should be mentioned as part of the submission process and the relevant declaration thereof should be included in the manuscript. (Perhaps as a subtitled section at the end of the methods section.)

2. At least one peer reviewer should be tasked with comparing the preregistration with the final manuscript using *a brief checklist that should also perfectly match the required checklist completed by the author in the preregistration document*. This would include the specification of the primary outcome, sample size justification, identity of the experimental conditions, analytic approach, and inclusion/exclusion criteria.
3. Any unmentioned discrepancies between the authors' rendition and the completed manuscript should either result in a rejection of the manuscript or their inclusion being made a condition of acceptance.
4. Readers should be brought into the process and rewarded with the authorship of a no-charge, published "errata discovery" or "addendum" of some sort since these could constitute fail-safe candidates for both checking and enforcing this policy. (This might eventually become common enough to expel the personal onus editors seem to associate with publishing errata or negative comments.)

Now, of course, these suggestions entail some extra expense from the journals' perspective but academic publishers can definitely afford to hire an extra staff member or, heaven forbid, even pay a peer reviewer an honorarium when tasked with comparing the preregistered protocol with the manuscript he or she is reviewing. (After all, publishers can always fall back on one of their chief strengths, which is to pass on any additional costs to authors and their institutions.)

On a positive note, there is some evidence that compliance with preregistration edicts may have begun to improve in the past decade or so—at least in some disciplines. Kaplan and Irvin (2015), for example, conducted a natural experiment in which large (defined as requiring more than $500,000 in direct costs) National Heart, Lung, and Blood Institute-funded cardiovascular RCTs were compared before and after preregistration was mandated by clinicaltrials.gov.

Unlike preregistration requirements announced by journals or professional organizations, this one has apparently been rigorously enforced since Kaplan and Irvin found that 100% of the located 55 trials published after

2000 were registered as compared to 0% prior thereto. (Note the sharp contrast to other disciplines without this degree of oversight, as witnessed by one disheartening study [Cybulski, Mayo-Wilson, & Grant, 2016] which found that, of 165 health-related *psychological* RCTs published in 2013, only 25 [15%) were preregistered.)

Perhaps equally surprisingly (and equally heartening), the 2015 Kaplan and Irwin cardiovascular study also found a precipitous drop in publication bias, with trials published prior to 2000 reporting "significant benefit for their primary outcome" in 17 of 30 (57%) studies versus 8% (or 2 of 25) after 2000 ($p < 0.0005$). The authors attributed this precipitous drop in positive findings to one key preregistration requirement of the ClinicalTrials.gov initiative.

> Following the implementation of ClinicalTrials.gov, investigators were required to prospectively declare their primary and secondary outcome variables. Prior to 2000, investigators had a greater opportunity to measure a range of variables and to select the most successful outcomes when reporting their results. (p. 8)

Now while such a dramatic drop in positive results may not be particularly good news for sufferers of heart disease, their families, or congressional oversight committees, it constitutes a large step back from the rampant publication bias discussed earlier and hopefully a major step forward for the reproducibility movement for several reasons.

1. In mature disciplines such as cardiovascular research, dramatic new findings *should* decrease over time as the availability of low-hanging fruit decreases.
2. Clinical RCTs are expected to adhere to the CONSORT agreement, and strict adherence thereto precludes the false-positive producing effects of most (if not all) of the QRPs listed in Chapter 3.
3. Well-funded National Institutes of Health (NIH) clinical trials are also associated with adequate sample sizes—hence adequate statistical power—which is not the case in the majority of social science experiments.

And lest it appear that social science experiments are being ignored here with respect to changes from preregistration to publication, the National Science Foundation (NSF)-sponsored Time-sharing Experiments for the

Social Sciences (TESS) provides an extremely rare opportunity for comparing preregistered *results* with published *results* for a specialized genre of social science experiments. The program itself involved embedding "small" unobtrusive interventions (e.g., the addition of a visual stimulus or changes in the wording of questions) into national surveys conducted for other purposes. Franco, Malhotra, and Simonovits (2014) were then able to compare the unpublished results of these experiments with their published counterparts since the NSF required not only the experimental protocols and accruing data to be archive prior to publication, but also the *study* results as well.

The authors were able to locate 32 of these studies that had been subsequently published. They found that (a) 70% of the published studies did not report all the outcome variables included in the protocol and (b) 40% did not report all of the proposed experimental conditions. Now while this could have been rationalized as editorial pressure to shorten the published journal articles, another study by the Franco et al. team discovered a relatively unique wrinkle to add to the huge publication bias and QRP literatures: "Roughly two thirds of the reported tests [were] significant at the 5% level compared to about one quarter of the unreported tests" (p. 10). A similar result was reported for political science studies drawn from the same archive (Franco, Malhotra, & Simonovits, 2017).

And if anyone needs to be reminded that registry requirements alone aren't sufficient, one extant registry has even more teeth than ClinicalTrials.gov. This particular registry is unique in the sense that it potentially controls access to hundreds of billions in profit to powerful corporations. It is also an example of a registry developed by a government agency that closely examines and evaluates all submitted preregistrations before the applicants can proceed with their studies, as well as the results after the trials are completed.

That honor goes to the US Food and Drug Administration (FDA). The FDA requires that positive evidence of efficacy (in the form of randomized placebo or active comparator trials) must be deposited in its registry and that this proposed evidence must be evaluated by its staff before a specific medical condition can be approved for a specific diagnosis.

To fulfill this responsibility, the agency's registry requires a prospective protocol, including the analysis plan, the actual RCT data produced, and the results thereof in support of an application for either marketing approval or a change in a drug's labeling use(s). FDA statisticians and researchers

then review this information to decide whether the evidence is strong enough to warrant approval of each marketing application. Such a process, if implemented properly, should preclude the presence of a number of the QRPs described in Chapter 3.

Alas, what the FDA does not review (or regulate) are the *published* results of these trials that wind up in the peer reviewed scientific literature. Nor does it check those results against what is reported in its registry. But what if someone else did?

Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy

Erick Turner, Annette Matthews, Eftihia Linardatos, et al. (2008)

The authors of this study took advantage of available information from the FDA approval process to evaluate the possibility of bias in pharmaceutical industry-funded research. The studies under review were all randomized, placebo-controlled evaluations of 12 antidepressant agents approved by the FDA between 1987 and 2004.

Of the 74 FDA-registered studies, 38 were judged by the FDA as being positive and 37 (97%) of these were published. Of the 36 trials that were judged to be *negative* (n = 24) or *questionable* (n = 12), only 14 (39%) wound up being published.

So far all we have is more unnecessary evidence of publication bias since positive results were almost two and a half times more likely to be published. The next finding, however, is what should be of primary interest to us here.

Of the FDA-judged positive studies published, all (100%) were presented as positive in the final report. (No surprise here since pharmaceutical companies aren't likely to suppress good news.) However of the 12 questionable studies, the 6 that were published all appeared in the published record as *positive* (i.e., a miraculous 50% improvement in effectiveness between the time the FDA reviewed the trial results and the time they appeared in print). Of the 8 published negative trials, five (64%) of the ineffective drugs somehow, magically, became effective as a function of the publication process.

Other, equally disturbing finding, reported were

> The methods reported in 11 journal articles appeared to depart from the pre-specified methods reflected in the FDA reviews. . . . Although for each of these studies the finding with respect to the protocol-specified primary outcome was non-significant, each publication highlighted a positive results as if it were the primary outcome. [Sound familiar by now?] The non-significant results of the pre-specified primary outcomes were either subordinated to non-primary positive results (in two reports) or omitted (in nine). (p. 255)

And,

> By altering the apparent risk–benefit ratio of drugs, selective publication can lead doctors to make inappropriate prescribing decisions that may not be in the best interest of their patients and, thus, the public health. (p. 259)

Another team conducted a study around the same time period (Rising, Bacchetti, & Bero, 2008) employing different FDA studies and reported similar results—including the propitious changes from registry to journal articles. However, in 2012, Erick Turner (with Knowepflmacher & Shapey) basically repeated his 2008 study employing antipsychotic trials and found similar (but less dramatic) biases—hopefully due to Erick's alerting pharmaceutical companies that someone was watching them, but more likely due to the greater efficacy of antipsychotic drugs. (Or perhaps placebos are simply less effective for patients experiencing psychotic symptoms than in those with depression.)

One final study involving an often overlooked preregistration registry: we don't conceptualize them in this way, but federally mandated institutional review boards (IRBs) and institutional animal care and use committees (IACUCs) are registries that also require protocols containing much of the same basic information required in a preregistration.

The huge advantage of these local regulatory "registries" is that it is illegal for any institution (at least any that receive federal funding) to allow the recruitment of participants (human or animal) for any research purposes without first submitting such a protocol for approval by a committee designated for this purpose. More importantly most of these institutions are quite conscientious in enforcing this requirement since federal research funding

may be cut off for violations. So, obviously, such registries would constitute an excellent opportunity for comparing regulatory protocols with their published counterparts, but for some inexplicable reason IRB and IACUC records are considered proprietary.

However, occasionally investigators are provided access to selected IRBs for research purposes. Chan, Hrobjartsson, Haahr, and colleagues (2004), for example, were able to obtain permission from two Danish IRBs to identify 102 experimental protocols submitted between 1994 and 1995 that had subsequently been published. Each proposed protocol was then compared to its published counterpart to identify potential discrepancies in the treatment or the specified primary outcome. The identified changes from application to publication in the pre-specified primary outcomes were that some (a) magically became secondary outcomes, (b) were replaced by a secondary outcomes, (c) disappeared entirely, or (d) regardless of status, the outcomes used in the power calculations required by the IRB protocols differed from those reported in the published articles (which were necessitated by the previous three changes).

Of the 102 trials, 82 specified a primary outcome (it is "puzzling" that 20 did not). Of these, 51 (62%) had made at least one of the four just-mentioned changes. And, not surprisingly, the investigators found that

> The odds of a particular outcome being fully reported were more than twice as high if that outcome was statistically significant. Although the response rate was relatively low, one of the most interesting facets of this study was a survey sent to the studies' authors. Of the 49 responses received, 42 (86%) actually "denied the existence of unreported outcomes despite clear evidence to the contrary." (p. 2457)

So while we don't conceptualize them as such, federally mandated IRBs and IACUCs are actually preregistration repositories for all studies involving human or animal participants that basically contain all of the necessary information required for preregistered protocols. (Sometimes a proposed study is granted exemption from IRB review, but the reasons for such decisions are also on file.)

While these committees differ among institutions from being overly officious to operating with a wink and a nod, there is no reason why all proposals couldn't (a) collect the requisite information on the same standardized checklist form suggested earlier and (b) register that information

on either their institutional websites or a national registry. Obviously, such a process would elicit massive hues and cries from everyone—investigators, administrators, and their lobbyists citing both financial and privacy issues—but so what?

After all, shouldn't an IRB that is instituted for research participants' protection also protect those participants from squandering their time and effort on the incessant production of fallacious research results? Also, since no participant identifiers are contained in an IRB application, no privacy concerns can emanate from them except for study investigators. And anonymity is the last thing investigators want since they gladly affix their names to the resulting published articles.

While the following three potential objections may smack of once again raising strawmen, they probably need to be addressed anyway.

1. *Cost*: Admittedly in research-intensive institutions the IRB-IACUC regulatory process would require at least one additional full-time staff member to upload all approved proposals and attached amendments (or at the very least the completed standardized checklist just suggested) to a central registry. (Or perhaps someone could write a program to do so automatically.) Just as obviously, some effort (preferably automated) will be necessary to ensure that none of the required fields is empty. However, research institutions receive very generous indirect costs (often exceeding 50% of the actual research budget itself) so there is adequate funding for an integral research function such as this.
2. *Release date*: The proposal (or possibly simply the minimal information containing hypotheses, primary outcomes, experimental conditions, sample size, study design, and analytic approach), while already uploaded and dated, could be released only upon submission of the final manuscript for publication and only then following the journal's decision to submit it to the peer review process. This release process could be tightened by the principal investigator granting access to the proposal (and its amendments) only to the journal to which it is submitted. However, any such restrictions would be lifted once the manuscript had been published or after a specified period of time. (Both the submitted manuscript and the published article would have to include a direct link to the archived IRB proposal or checklist.)
3. *Amendments to regulatory proposals*: As mentioned, IRBs and IACUCs differ significantly in their degree of oversight and conscientiousness.

> My familiarity with IRBs extends only to those representing academic medical centers, which are probably more rigorous than their liberal arts counterparts. However, any serious IRB or IACUC should require dated amendments to a proposal detailing changes in (a) sample size (increased or decreased) along with justifications thereof, (b) experimental conditions (including changes to existing ones, additions, and/or deletions), (c) primary outcomes, and (d) analytic approaches. All such amendments should be attached to the original protocol in the same file, which would certainly "encourage" investigators to include any such changes in their manuscripts submitted for publication since their protocols would be open to professional scrutiny.

The same procedures should be implemented by funding agencies. All federal (and presumably most private) funders require final reports which include the inferential results obtained. While these are seldom compared to the original proposal prior to simply being checked off and filed away by bureaucrats (probably often without even being read), they, too, should be open to public and professional scrutiny.

A final note: preregistration of protocols need not be an onerous or time-consuming process. It could consist of a simple six- or seven-item checklist in which each item requires no more than a one- or two-sentence explication for (a) the primary hypothesis and (b) a justification for the number of participants to be recruited based on the hypothesized effect size, the study design, and the resulting statistical power emanating from them. The briefer and less onerous the information required, the more likely the preregistration process and its checking will be implemented.

Data Sharing

Universal preregistration of study protocols and their comparison to published manuscripts may be the single most effective externally imposed strategy for decreasing the presence of false-positive results. However, the universal requirement for sharing research data isn't far behind.

While data sharing is important in all fields, the main emphasis here will be on data generated in the conduct of discrete experimental or correlational research studies rather than the mammoth databases produced

in such fields as genomics, physics, or astronomy which tend to be already archived, reasonably well-documented, and available for analysis. Also excluded are large-scale surveys or datasets prepared by the government or major foundations. (Naturally it is important that all code, selection decisions, variable transformations, and analytic approaches used in any published works be available in archived preregistrations for any typed of empirical study.)

Donoho, Maleki, Shahram, and colleagues (2009) make the argument that "reproducible computational research, in which all details of computations—code and data—are made conveniently available to others, is a necessary response to this credibility crisis" (p. 8). The authors go on to discuss software and other strategies for doing so and debunk excuses for not doing so. Or, as Joelle Pineau observed regarding artificial intelligence (AI) research (Gibney, 2019): "It's easy to imagine why scientific studies of the natural world might be hard to reproduce. But why are some algorithms irreproducible?" (p. 14) Accordingly, at the December 2019 Conference on Neural Information Processing Systems (a major professional AI meeting) the organizing committee asked participants to provide code along with their submitted papers followed by "a competition that challenged researchers to recreate each other's work." As of this writing, the results are not in, but the very attempt augurs well for the continuance and expansion of the reproducibility movement both for relatively new avenues of inquiry and across the myriad classic disciplines that comprise the scientific enterprise. Perhaps efforts such as this provide some hope that the "multiple-study" replication initiatives discussed in Chapter 7 will continue into the future.

At first glance data sharing may seem more like a generous professional gesture than a reproducibility necessity. However, when data are reanalyzed by separate parties as a formal analytic replication (aka *analytic reproduction*), a surprisingly high prevalence of errors favoring positive results occur, as discussed in Chapter 3 under QRP 6 (sloppy statistical analyses and erroneous results in the reporting of p-values).

So, if nothing else, investigators who are required to share their data will be more likely to take steps to ensure their analyses are accurate. Wicherts, Bakker, and Molenaar (2011), for example, found significantly more erroneously reported p-values among investigators who refused to share than those who did.

So, the most likely reasons for those who refuse to share their data are

1. The data were not saved,
2. The data were saved but were insufficiently (or too idiosyncratically or not at all) documented and hence could not be reconstructed by the original investigators or statisticians due to the previously discussed inevitable memory loss occurring with time,
3. The investigators planned to perform and publish other analyses (a process that should have also been preregistered and begun by the time the original article was published, although a reasonable embargo period could be specified therein),
4. Proprietary motives, in which case the study probably shouldn't have been published in the first place, and
5. Several other motives, some of which shouldn't be mentioned in polite scientific discourse.

However, data sharing, unlike preregistration, possesses a few nuances that deserve mention. First of all, the creation of interesting data often requires a great deal of time and effort, so perhaps it is understandable that an investigator resents turning over something so precious to someone else to download and analyze in search of a publication. Perhaps this is one reason that Gary King (1995) suggested that data creation should be recognized by promotion and tenure committees as a significant scientific contribution in and of itself. In addition, as a professional courtesy, the individual who created the data in the first place could be offered an authorship on a publication if he or she provides any additional assistance that warranted this step. But barring that, data creators should definitely be acknowledged and cited in any publication involving their data.

Second, depending on the discipline and the scope of the project, many principal investigators turn their data entry, documentation, and analysis over to someone else who may employ idiosyncratic coding and labeling conventions. However, it is the principal investigator's responsibility to ensure that the data and their labeling are clear and explicit. In addition, code for all analyses, variable transformations, and annotated labels, along with the data themselves, should be included in a downloadable file along with all data cleaning or outlier decisions. All of these tasks are standard operating

procedures for any competent empirical study, but, as will be discussed shortly, compliance is far from perfect.

The good news is that the presence of archived data in computational research sharing has increased since the 2011–2012 awakening, undoubtedly facilitated by an increasing tendency for journals to delineate policies to facilitate the practice. However, Houtkoop, Wagenmakers, Chambers, and colleagues (2018) concluded, based on a survey of 600 psychologists, that "despite its potential to accelerate progress in psychological science, public data sharing remains relatively uncommon." They consequently suggest that "strong encouragement from institutions, journals, and funders will be particularly effective in overcoming these barriers, in combination with educational materials that demonstrate where and how data can be shared effectively" (p. 70).

There are, in fact, simply too many important advantages of the data sharing process for it to be ignored. These include

1. It permits exact replication of the statistical analyses of published studies' as well as a resource to pilot new hypotheses; since data on humans and animals are both difficult and expensive to generate, we should glean as much information from them as possible.

And, relatedly,

2. It permits creative secondary analyses (including combining different datasets with common variables) to be performed if identified as such, thereby possibly increasing the utility of archived data.

But for those who are unmoved by altruistic motives, scientific norms are changing and professional requests to share data will continue to increase to the point where, in the very near future, failing to comply with those requests will actually become injurious to a scientist's reputation. Or, failing that, it will at the very least be an increasingly common publication requirement in most respectable empirical journals.

However, requirements and cultural expectations go only so far. As Robert Burns put it, "The best laid schemes o' mice an' men gang aft agley," or, more mundanely translated to scientific practice by an unknown pundit, "even the most excellent of standards and requirements are close to worthless in the absence of strict enforcement".

"Close" but not completely worthless as illustrated by a study conducted by Alsheikh-Ali, Qureshi, Al-Mallah, and Ioannidis (2011), in which the

first 10 original research papers of 2009 published in 50 of the highest impact scientific journals (almost all of which were medicine- or life sciences-oriented) were reviewed with respect to their data sharing behaviors. Of the 500 reviewed articles, 351 papers (70%) were subject to a data availability policy of some sort. Of these, 208 (59%) were not completely in compliance with their journal's instructions. However, "*none of the 149 papers not subject to data availability policies made their full primary data publicly available* [emphasis added]" (p. 1).

So to be even moderately effective, the official "requirement for data registration" will not result in compliance unless the archiving of said data (a) precedes publication and (b) is checked by a journal representative (preferably by a statistician) to ensure adequate documentation and transparent code. And this common-sense generalization holds for funding agency requirements as well, as is disturbingly demonstrated by the following teeth-grating study.

A Funder-Imposed Data Publication Requirement Seldom Inspired Data Sharing

Jessica Couture, Rachael Blake, Gavin McDonald, and Colette Ward (2018)

The very specialized funder in this case involved the Exxon Valdez Oil Spill Trustee Council (EVOSTC), which funded 315 projects tasked with collecting ecological and environmental data from the 1989 disaster. The Couture et al. team reported that designated staff attempted to obtain data from the 315 projects (a funding requirement), which in turn resulted in 81 (26%) of the projects complying with some data. . . . No data at all were received from 74% of the funded entities (23% of whom did not reply to the request and 49% could not be contacted).

Unfortunately, although the EVOSTC reported funding hundreds of projects, "the success of this effort is unknown as the content of this collection has since been lost [although surely not *conveniently*]." But, conspiracy theories aside, the authors do make a case that a recovery rate of 26% is not unheard of, and, while this may sound completely implausible, unfortunately it appears to be supported by at least some empirical

evidence, as witnessed by the following studies reporting data availability rates (which in no way constitutes a comprehensive or systematic list).

1. Wollins (1962) reported that a graduate student requested the raw data from 37 studies reported in psychology journals. All but 5 responded, and 11 (30% of the total requests) complied. (Two of the 11 investigators who did comply demanded control of anything published using their data, so 24% might be considered a more practical measure of compliance.)
2. Wicherts, Borsboom, Kats, and Molenaar (2006) received 26% of their requested 249 datasets from 141 articles published in American Psychiatric Association journals.
3. Using a very small sample, Savage and Vickers (2009) requested 10 datasets from articles published in *PLoS Medicine* and *PLoS Clinical Trials* and received only 1 (10%). This even after reminding the original investigators that both journals explicitly required data sharing by all authors.
4. Vines, Albert, Andrew, and colleagues (2014) requested 516 datasets from a very specialized area (morphological plant and animal data analyzed via discriminant analysis) and received a response rate of 19% (101 actual datasets). A unique facet of this study, in addition to its size, was the wide time period (1991 to 2011) in which the studies were published. This allowed the investigators to estimate the odds of a dataset becoming unavailable over time, which turned out to be a disappearance rate of 17% per year.
5. Chang and Li (2015) attempted to replicate 61 papers that did not employ confidential data in "13 well-regarded economics journals using author-provided replication files that include both data and code." They were able to obtain 40 (66%) of the requisite files. However, even with the help of the original authors, the investigators were able to reproduce the results of fewer than half of those obtained. However, data sharing was approximately twice as high for those journals that required it as for those that did not (83% vs. 42%).
6. Stodden, Seiler, and Ma (2018) randomly selected 204 articles in *Science* to evaluate its 2011 data sharing policy, of which 24 provided access information in the published article. Emails were sent to the remaining authors, of which 65 provided some data and/or

> code, resulting in a total of 89 (24 + 65) articles that shared at least some of what was requested. This constituted a 44% retrieval rate, the highest compliance rate of any reviewed here, and (hopefully not coincidentally) it happened to be the most recent article. From these 89 sets of data, the investigators judged 56 papers to be "potentially computationally reproducible," and from this group they randomly selected 22 to actually replicate. All but one appeared to replicate, hence the authors estimated that 26% of the total sample may have been replicable. [Note the somewhat eerie but completely coincidental recurrence of this 26% figure.]

All six sets of authors provided suggestions for improvement, especially around the adequacy of runnable source code. This is underlined in a survey of 100 papers published in *Bioinformatics* (Hothorn & Leisch, 2011), which found that adequate code for simulation studies was "limited," although what is most interesting about this paper is that the first author serves (or served) as the "reproducible research editor" for a biometric journal in which one of his task was to check the code to make sure it ran—a role and process which should be implemented by other journals. (Or perhaps alternately a startup company could offer this role on a per-article fee basis.)

Despite differences in methodologies, however, all of these authors would probably agree with Stodden and her colleagues' conclusion regarding *Science*'s data sharing guidelines (and perhaps other journal-unenforced edicts as well).

> Due to the gaps in compliance and the apparent author confusion regarding the policy, we conclude that, although it is a step in the right direction, this policy is insufficient to fully achieve the goal of computational reproducibility. Instead, we recommend that the journal verify deposit of relevant artifacts as a condition of publication. (p. 2588)

A disappointment, perhaps due to an earlier finding by the Stodden team (2013) which found a 16% increase in data policies occurring between 2011 and 2012 and a striking 30% increase in code policies in a survey of 170 statistically and computationally oriented journals.

But does universal access to data affect reproducibility? No one knows, but, speculatively, it should. If data are properly archived, transparently

documented, and complete (including variables not presented in the official publication and all variable transformations), the availability of such data may present a rare roadblock on Robert Park's path (straight, not "forked") from "foolishness to fraud."

How? By making it possible for other scientists or statisticians to ascertain (a) if more appropriate alternate analyses produce the same result and (b) if certain QRPs appeared to be committed. The process would be exponentially more effective if the universal requirement for data archiving could be coupled with the preregistration of study hypotheses and methods. And that may be in the process of occurring.

At the very least, journals should require all empirical research data to be permanently archived. It is not sufficient to "require" investigators to share their data upon request because too many untoward events (e.g., multiple job and computer changes, retirement, and dementia) can occur over time to subvert that process, even for investigators with the most altruistic of motives.

Journals should also not accept a paper until the archived data are checked for completeness and usability. Furthermore, a significant amount of money should be held in escrow by funders until the archived data are also checked. In one or both cases, an independent statistician should be designated to check the data, code, and other relevant aspects of the process as well as personally sign off on the end product of said examination. To facilitate this, the archived code should be written in a commonly employed language (a free system such as R is probably preferable, but software choices could be left up to investigators as long as they are not too esoteric). Everything should also be set up in such a way that all results can be run with a mouse click or two.

While short shrift has admittedly been given to purely computational research and reproducibility, this field is an excellent resource for suggestions concerning computational reproducibility. Sandve, Nekrutenko, Taylor, and Hovig (2013), for example, begin by arguing that ensuring the reproducibility of findings is as much in the self-interest of the original investigators as it is to the interests of others.

> Making reproducibility of your work by peers a realistic possibility sends a strong signal of quality, trustworthiness, and transparency. This could increase the quality and speed of the reviewing process on your work, the chances of your work getting published, and the chances of your work being taken further and cited by other researchers after publication. (p. 3)

They then offer 10 rules for improving the reproducibility of computational research that are largely applicable to the replication of all empirical research (and especially to the issue of sharing reproducible data). The explanations for these rules are too detailed to present here, and some are applicable only to purely computational efforts, hence, hopefully, I will be forgiven for my rather abrupt abridgments and selections. Hopefully also, the original authors will excuse the verbatim repetition of the rules themselves (all quoted passages were obtained from pp. 2–3), which apply to all types of complex datasets).

Rule 1: For every result, keep track of how it was produced. This basically reduces to maintaining a detailed analytic workflow. Or, in the authors' words: "As a minimum, you should at least record sufficient details on programs, parameters, and manual procedures to allow yourself, in a year or so, to approximately reproduce the results."

Rule 2: Avoid manual data manipulation steps. In other words, use programs and codes to recode and combine variables rather than perform even simple data manipulations *manually*.

Rule 3: Archive the exact versions of all external programs used. Some programs change enough over time to make exact replication almost impossible.

Rule 4: Version control all custom scripts. Quite frankly, this one is beyond my expertise so for those interested, the original authors suggest using a "version control system such as Subversion, Git, or Mercurial." Or, as a minimum, keep a record of the various states the code has taken during its development.

Rule 5: Record all intermediate results, when possible in standardized formats. Among other points, the authors note that "in practice, having easily accessible intermediate results may be of great value. Quickly browsing through intermediate results can reveal discrepancies toward what is assumed, and can in this way uncover bugs or faulty interpretations that are not apparent in the final results."

Rule 7: Always store raw data behind plots. "As a minimum, one should note which data formed the basis of a given plot and how this data could be reconstructed."

Rule 10: Provide public access to scripts, runs, and results.

Of course, while all of these rules may not be completely applicable for a straightforward experiment involving only a few variables, their explication does illustrate that the process of data archiving can be considerably more complicated (and definitely more work intensive) than is generally perceived by anyone who hasn't been involved in the process. Most journals will require only the data mentioned in a published article to be archived, and that may be reasonable. But again, someone at the journal- or investigator-level must at least rerun all of the reported analyses using only the archived information to ensure that the same results are identically reproduced for tables, figures, and text. Otherwise it is quite improbable that said results will be reliably reproducible.

Of course, sharing one's analytic file doesn't guarantee that the final analysis plan wasn't arrived at via multiple analyses designed to locate the one resulting in the most propitious p-value. As a number of statisticians note, and both Steegen, Tuerlinckx, Gelman, and Vanpaemel (2016) and our favorite team of Simonsohn, Simmons, and Nelson (2015) empirically illustrate, different analytic decisions often result in completely different inferential results. And while such decisions are capable of being quite reasonably justified a posteriori, one of our authors has previously reminded us that in a "garden of forking paths, whatever route you take seems predetermined."

Ironically, the Steegen et al. team illustrated this potential for selectively choosing an analytic approach capable of producing a statistically significant p-value by using the study employed by Andrew Gelman to illustrate his garden of forking paths warning. (The study—and a successful self-replication thereof—it will be recalled was conducted by Durante, Rae, and Griskevicius (2013) who "found" that women's fertility status was influenced both their religiosity and political attitudes.) The Steegen team's approach involved

1. Employing the single statistical result reported in the Durante et al. study,
2. Constructing what they termed the "data multiverse," which basically comprised all of the *reasonable* coding and transformation decisions possible (120 possibilities in the first study and 210 in the replication), and then
3. Running all of these analyses and comparing the p-values obtained to those in the published article.

The authors concluded that investigator data processing choices are capable of having a major impact on whether or not significant p-values are obtained in an observational or experimental dataset. For the studies employed in their demonstration the authors suggest that

> One should reserve judgment and acknowledge that the data are not strong enough to draw a conclusion on the effect of fertility. The real conclusion of the multiverse analysis is that there is a gaping hole in theory or in measurement, and that researchers interested in studying the effect of fertility should work hard to *deflate* the multiverse. The multiverse analysis gives useful directions in this regard. (p. 708)

The Simonsohn team (whose work which actually preceded this study) arrived at the same basic conclusions and provided, as is their wont, a statistical approach ("specification-curve analysis") for evaluating the multiple results obtained from these multiple, defensible, analytic approaches.

Both the Steegen et al. (2016) and the Simonsohn et al. (2015) articles demonstrate that different analytic approaches are in some cases capable of producing both statistically significant and non-significant results. And certainly some investigators may well analyze and reanalyze their data in the hope that they will find an approach that gives them a p-value < 0.05—thereby suggesting the need for a new QRP designation to add to our list or simply providing yet another example of p-hacking.

However, the recommendation that investigators analyze and report all of the "reasonable scenarios" (Steegen et al., 2016) or "multiple, defensible, analytic approaches" (Simonsohn et al., 2015) is, in my opinion, probably going a bridge too far. Especially since the first set of authors found an average of 165 possible analyses in a relatively simple study and its replication. So perhaps the group should have stuck with their advice given in their iconic 2011 article which involved (a) the preregistration of study analysis plans and (b) reporting results without the use of covariates.

Materials Sharing

Types of experimental materials vary dramatically in form and portability from discipline to discipline. In the types of psychology experiments

replicated via the Open Science Collaboration or the Many Labs initiative these range from questionnaires to graphic-laden scenarios along with explicit (hopefully) standardized instruction to participants and research assistants. These are easily shared via email, Drop Box, or some other internet delivery system, and compliance with requests for such information doesn't appear to be especially problematic, as witnessed by the degree of cooperation afforded to the above-mentioned multistudy replication initiatives.

Other setups in other disciplines can be more idiosyncratic, but it is difficult to imagine that many requests to visit a laboratory to inspect equipment or other apparatuses would be denied to interested investigators. For most bench research, the equipment used is often fairly standard and purchasable from a commercial supplier if the replicating lab doesn't already possess it. For studies employing sophisticated and expensive equipment, such as the functional magnetic resonance imaging (fMRI) studies described earlier, investigators should include all of the relevant information in their published documents or in an accessible archive.

In those instances where cooperation is not provided, the published results should be (and are) viewed with the same suspicion by the scientific community as afforded to unpublished discovery claims. The cold fusion debacle constituted an unusual example of this in the sense that the specifications for the apparatus were apparently shared but not the procedures employed to generate the now infamous irreproducible results.

Material sharing 2.0: Timothy Clark (2017), an experimental biologist, suggests taking the sharing of materials and procedures a step farther in a single-page *Nature* article entitled "Science, Lies and Video-Taped Experiments." Acknowledging the difficulties, he succinctly presents the following analogy: "If extreme athletes can use self-mounted cameras to record their wildest adventures during mountaintop blizzards, scientists have little excuse not to record what goes on in lab and field studies" (p. 139).

Perhaps his suggestion that journals should require such evidence to be registered (and even used in the peer review process) may *currently* be unrealistic. But the process would certainly (a) facilitate replication, (b) serve as an impressive and time-saving teaching strategy, and (c) discourage scientific misconduct. And there is even a journal partly designed to encourage the process (i.e., the *Journal of Visualized Experiments*.)

And Optimistically: A Most Creative Incentive for Preregistration and Data/Material Sharing

At first glance the reinforcement strategy for principled research practice about to be discussed may smack of the simplistic stickers used as reinforcements for preschoolers. However, the Open Science Network should never be underestimated, as witnessed by its advocacy of "badges" awarded to published studies for which the investigator has pledged (a) the availability of data, (b) experimental materials (which, as just mentioned, can obviously vary quite dramatically from discipline to discipline), and/or (c) an archived preregistered protocol.

The potential operative components of this simple reinforcement strategy are that paper badges affixed to a research article may potentially

1. Encourage other scientists to read the accompanying article since its results promise to be more reproducible, citable, and perhaps even more important;
2. Encourage colleagues and scientists interested in conducting secondary analyses of data or performing replications to not only read and cite the article but possibly provide its author(s) with collaborative opportunities. Piwowar, Day, and Fridsma (2007), for example, found that the citation rate for 85 cancer microarray clinical trial publications which shared *usable* research data was significantly higher compared to similar studies which did not do so;
3. Identify the badged authors as principled, careful, modern scientists; and
4. Potentially even increase the likelihood of future acceptances in the journal in which the badged article appears.

Interestingly, an actual evaluation of the badge concept has been conducted (Kidwell, Ljiljana, Lazarević, et al., 2016). Beginning in January of 2014, the high-impact journal *Psychological Science* was somehow cajoled into providing its authors with "the opportunity to signal open data and materials if they qualified for badges that accompanied published articles." (The badges were supplied by the journal's editorial team if the authors so requested and gave some "reasonable" evidence that they were indeed meeting the criteria.)

The evaluation itself consisted of (a) a before-and-after comparison of reported offers for data sharing prior to that date, (b) a control comparison

of the percentage of such offers in four other high-impact psychological journals that did not offer badges, and (c) an assessment of the extent to which these offers corresponded to the actual availability of said data.

The results were quite encouraging. In brief,

1. From a baseline of 2.5%, *Psychological Science* reported open data sharing increased to an average of 22.8% of articles by the first half of 2015 (i.e., in slightly over 1 year after the advent of badges);
2. The four comparison journals, on the other hand, while similar at baseline to *Psychological Science*, averaged only 2.1% thereafter (i.e., as compared to 22.8% in *Psychological Science*);
3. With respect to actual availability of usable data, the results were equally (if not more) impressive, with the *Psychological Science* articles that earned badges significantly outperforming the comparison journals. Perhaps a more interesting effect, however, involved a comparison of the availability of *usable* data of *Psychological Science* articles announcing availability with badges versus the *Psychological Science* articles announcing availability but *without* badges. For the 64 *Psychological Science* articles reporting availability of data archived on a website or repository, 46 had requested and been awarded a data sharing badge while 18 had not. Of those with a badge, 100% actually had datasets available, 82.6% of which were complete. For those who announced the availability of their data but did not have a badge, 77.7% ($N = 14$) made their data available but only 38.9% ($N = 7$) of these had complete data. And, finally,
4. The effects of badges on the sharing of materials were in the same direction as data sharing, although not as dramatic.

Now granted, these numbers are relatively small and the evaluation itself was comparative rather than a randomized experiment, but the authors (who transparently noted their study's limitations) were undoubtedly justified in concluding that, "Badges are simple, effective signals to promote open practices and improve preservation of data and materials by using independent repositories" (p. 1).

And that, Dear Readers, concludes the substantive subject matter of this book, although the final chapter will present a few concluding thoughts.

References

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. (2011). Public availability of published research data in high-impact journals. *PLoS ONE, 6*, e24357.

Bem, D. J. (1987). Writing the empirical journal article. In M. Zanna & J. Darley (Eds.), *The compleat academic: A practical guide for the beginning social scientist* (pp. 171–201). Mahwah, NJ: Lawrence Erlbaum Associates.

Chan, A. W., Hrobjartsson, A., Haahr, M. T., et al. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association, 29*, 2457–2465.

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say "usually not." Finance and Economics Discussion Series. http://dx.doi.org/10.17016/FEDS.2015.083

Clark, T. D. (2017). Science, lies and video-taped experiments. *Nature, 542*, 139.

Couture, J. L., Blake, R. E., McDonald, G., & Ward, C. L. (2018). A funder-imposed data publication requirement seldom inspired data sharing. *PLoS ONE, 13*, e0199789.

Cybulski, L., Mayo-Wilson, E., & Grant, S. (2016). Improving transparency and reproducibility through registration: The status of intervention trials published in clinical psychology journals. *Journal of Consulting and Clinical Psychology, 84*, 753–767.

De Angelis, C. D., Drazen, J. M., Frizelle, F. A., et al. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine, 351*, 1250–1252.

Donoho, D. L., Maleki, A., Shahram, M., et al. (2009). Reproducibility research in computational harmonic analysis. *Computing in Science & Engineering*, *11*, 8–18.

Durante, K., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science, 24*, 1007–1016.

Ewart, R., Lausen, H., & Millian, N. (2009). Undisclosed changes in outcomes in randomized controlled trials: An observational study. *Annals of Family Medicine, 7*, 542–546.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science, 7*, 8–12.

Franco, A., Malhotra, N., & Simonovits, G. (2017). Underreporting in political science survey experiments: Comparing questionnaires to published results. *Political Analysis, 23*, 306–312.

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist, 102*, 460–465.

Gibney, E. (2019). This AI researcher is trying to ward off a reproducibility crisis. *Nature, 577*, 14.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Hothorn, T., & Leisch, F. (2011). Case studies in reproducibility. *Briefings in Bioinformatics, 12*, 288–300.

Houtkoop, B. L., Wagenmakers, E.-J., Chambers, C., et al. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science, 1*, 70–85.

Huić, M., Marušić, M., & Marušić, A. (2011). Completeness and changes in registered data and reporting bias of randomized controlled trials in ICMJE journals after trial registration policy. *PLoS ONE, 6*, e25258.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS ONE, 10*, e0132382.

Kerr, N. L. (1991). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Kerr, N. L., & Harris, S. E. (1998). *HARKing-hypothesizing after the results are known: Views from three disciplines*. Unpublished manuscript. Michigan State University, East Lansing (not obtained).

Kidwell, M. C., Ljiljana B. Lazarević, L. B., et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*, e1002456.

King, G. (1995). Replication, replication. *PS: Political Science and Politics, 28*, 443–499.

Lin, W., & Green, D. P. (2016). Standard operating procedures: A safety net for pre-analysis plans. *Political Science and Politics, 49*, 495–500.

Mathieu, S., Boutron, I., Moher, D., et al. (2009). Comparison of registered and published primary outcomes in randomized controlled trials. *Journal of the American Medical Association, 302*, 977–984.

Mills, J. L. (1993). Data torturing. *New England Journal of Medicine, 329*, 1196–1199.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600–2606.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE, 2,* e308.

Rising, K., Bacchetti, P, & Bero, L. (2008). Reporting bias in drug trials submitted to the Food and Drug Administration: Review of publication and presentation. *PLoS Medicine, 5*, e217.

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE, 4*, e7078.

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology, 9*, e1003285.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. Manuscript available at http://ssrn.com/abstract=2694998

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science 11*, 702–712.

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE, 8,* e67111.

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences, 115*, 2584–2589.

Turner, E. H., Knowepflmacher, D., & Shapey, L. (2012). Publication bias in antipsychotic trials: An analysis of efficacy comparing the published literature to the US Food and Drug Administration database. *PLoS Medicine, 9*, e1001189.

Turner, E. H., Matthews, A. M., Linardatos, E., et al. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252–260.

Vines, T. H., Albert, A. Y. K., Andrew, R. L., et al. (2014). The availability of research data declines rapidly with article age. *Current Biology, 24*, 94–97.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6,* e26828.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726–728.

Wollins, L. (1962). Responsibility for raw data. *American Psychologist, 17*, 657–658.

11

A (Very) Few Concluding Thoughts

So where does the reproducibility initiative go from here? Certainly no one knows, although the easy answer is for reproducibility methodologists and advocates to continue their past and current efforts and continue the process of mentoring the next generation of researchers in their image, as well as including reproducibility concepts in all of their teaching activities (formal and informal).

Unfortunately, a precept in one of the several disciplines in which I have worked (preventive and health-seeking behaviors) was that "knowledge alone is inadequate to ensure the adoption of salutary behaviors and the avoidance of harmful ones."

However, this precept was not meant to imply that health education is worthless—only that the resulting knowledge is an insufficient condition for the adoption of such behaviors. So, in somewhat of a generalization leap, why shouldn't this precept also apply to increasing the reproducibility of scientific findings?

Educational Interventions

In the present context, the preventive behaviors in most dire need of adoption involve the responsible conduct of reproducible research. Hence, if nothing else, all students in all disciplines need to be taught the key importance of recognizing and avoiding questionable research practices (QRPs) along with the basics (and importance) of the replication process.

There are now many excellent resources available to anyone charged with this type of instruction. One resides in accessible articles on the topic which haven't been discussed previously such as those by Asendorpf, Conner, De Fruyt, and colleagues (2013) and by Munafò, Nosek, Bishop, and colleagues (2017)—both of which provide excellent overviews of the reproducibility process and suggestions for teaching its precepts to students and implementing them in practice.

The Problem with Science. R. Barker Bausell, Oxford University Press (2021). © Oxford University Press.
DOI: 10.1093/oso/9780197536537.003.0012

With respect to the educational process, Munafò et al. suggest (in addition to a formal course, which is more common in the social sciences than their life and physical counterparts) that

> The most effective solutions [for both students and faculty] may be to develop educational resources that are accessible, easy-to-digest . . . web-based modules for specific topics, and combinations of modules that are customized for particular research applications). A modular approach simplifies the process of iterative updating of those materials. Demonstration software and hands-on examples may also make the lessons and implications particularly tangible to researchers at any career stage . . . [such as] the Experimental Design Assistant (https://eda.nc3rs.org.uk) supports research design for whole animal experiments, while *P*-hacker (http://shinyapps. org/apps/p-hacker/) shows just how easy it is to generate apparently statistically significant findings by exploiting analytic flexibility. (p. 2)

While both the Asendorpf et al. (2013) and the Munafò et al. (2017) articles are too comprehensive to abstract here and their behavioral dicta have been discussed previously, each deserves to be read in its entirety. However, in buttressing the argument that methodological and statistical resources should be sought after by (and available to) all *investigators*, the latter recommends a model instituted by the CHDI Foundation (which specializes in research on Huntington's disease). Here, a committee of independent statisticians and methodologists are available to offer "a number of services, including (but not limited to) provision of expert assistance in developing protocols and statistical analysis plans, and evaluation of prepared study protocols" (p. 4).

Of course, students can and should be brought into such a process as well. In the conduct of science, few would argue that hands-on experience is one of the, if the not the most, effective ways to learn how to *do* science. Hence the Munafò et al. paper describes a resource designed to facilitate this process in psychology available under the Open Science framework umbrella called the Collaborative Replications and Education Project (https://osf.io/wfc6u/), in which

> A coordinating team identifies recently published research that could be replicated in the context of a semester-long undergraduate course on research methods. A central commons provides the materials and guidance

> to incorporate the replications into projects or classes, and the data collected across sites are aggregated into manuscripts for publication. (p. 2)

The Asendorpf paper also strongly advocates student replication projects and suggests what a reproducibility curriculum should emphasize, such as (a) the methodological basics of both single and multiple experimentation, (b) transparency, (c) best research practices (primarily the avoidance of QRPs), and (d) the introduction of students to online resources for preregistration of protocols and registries for archiving data. (A number of other investigators also advocate student replication projects; see Frank & Saxe, 2012; Grahe, Reifman, Hermann, et al., 2012; Jekel, Fiedler, Torras, et al., 2019—the latter summarizing how a student replication initiative [the *Hagen Cumulative Science Project*] can be practically implemented based on its track record, which, as of this writing, has produced 80+ student replications.)

In addition, regardless of discipline, all science students (undergraduates through postdocs) should be exposed to the Consolidated Standards of Reporting Trials (CONSORT; http://www.consort-statement.org/), along with one or more of its relevant 20+ specialized extensions (including one on social science experimentation). As mentioned previously, CONSORT is the gold standard of publishing guidelines, but a plethora of others exist for almost all types of research such as animal studies (ARRIVE), observational studies (STROBE), diagnostic studies (STARD), quality improvement efforts (SQUIRE), systematic reviews and meta-analyses (PRISMA), and so forth. All are accessible through EQUATOR (Enhancing the Quality and Transparency of Health Research: http://www.equator-network.org/), and, while most target health-related inquiry, a bit of translation makes them relevant to all empirical disciplines employing human or animal participants. Students should also be directed to the exemplary, extremely informative, huge (and therefore somewhat intimidating) Open Science Framework website (https://osf.io/), as well as the National Institutes of Health (NIH) Rigor and Reproducibility initiative (https://www.nih.gov/research-training/rigor-reproducibility).

Most importantly of all, however, mentorships should be formalized and supplemented because sometimes some of the most "successful" (i.e., sought after) research mentors in an institution may also be the most adept at obtaining $p < 0.05$ driven irreproducible results. For, as Andrew Gelman (2018) succinctly (as always) reminds us,

> The big problem in science is not cheaters or opportunists, but sincere researchers who have unfortunately been trained to think that every statistically "significant" result is notable.

And a big problem with un-training these "sincere" researchers involves how dearly we all hold on to something we have learned and internalized. So the best educational option is probably to keep the sanctity of obtaining p-values at all costs from being learned in the first place. If not by one-on-one mentoring, at least by providing graduate students, postdocs, and even advanced undergraduates with both conventional educational opportunities as well as online tutorials.

So What Is the Future of Reproducibility?

Burdened by the constantly regretted decision to employ Yogi Berra as a virtual mentor, I cannot in good faith answer this question directly. Instead, a pair of alternative scientific scenarios will be posited to allow others far wiser than I to choose the more likely of the two.

In one future, the reproducibility initiative will turn out to be little more than a publishing opportunity for methodologically oriented scientists—soon replaced by something else and forgotten by most—thereby allowing it to be reprised a few decades later under a different name by different academics.

In this future, publication bias will remain rampant and journals will continue to proliferate, as will the false-positive results they publish. After acknowledging the importance of reproducibility, most scientists will soon grow bored with its repetitively strident, time-consuming precepts and simply redouble their efforts in search of new publishing opportunities.

No harm done and few if any benefits achieved—at least as far as the social sciences are concerned, since they are cyclical rather than cumulative in nature anyway. And as for psychology, well, perhaps Paul Meehl's 1990 description of his discipline decades ago will continue to be prescient for future decades as well.

> Theories in the "soft areas" of psychology have a tendency to go through periods of initial enthusiasm leading to large amounts of empirical

> investigation with ambiguous over-all results. This period of infatuation is followed by various kinds of amendment and the proliferation of ad hoc hypotheses. Finally, in the long run, experimenters lose interest rather than deliberately discard a theory as clearly falsified. (p. 196)

In the other alternative future, the avoidance of QRPs will be added to the professional curricula to reduce the prevalence of false-positive results in all the major disciplines. Practicing researchers will become enculturated into avoiding problematic practices and rewarded for engaging in salutary ones—partly through peer pressure, partly because they have always wanted to advance their science by producing credible findings (but were stymied by conflicting cultural/economic forces), partly because committing QRPs has become increasing difficult due to publishing requirements and increased professional scrutiny, and partly because teaching is part of their job description and they have come to realize that there is nothing more important for their students and mentees to learn than the basic principles of conducting reproducible science.

But for those investigators who refuse or are unable to change? They'll gradually fade away and be replaced by others for whom acculturation into the reproducibility paradigm won't be an issue because it has already occurred. And the most conspicuous result of this evolution will be scientific literatures no longer dominated by positive results but replaced by less scintillating, negative studies suggesting more fruitful lines of inquiry.

So Which Alternative Future Is More Likely?

If Las Vegas took bets on the two alternative futures, the smart money would undoubtedly favor the first. After all, conducting reproducible research is monetarily more expensive and entails extra effort and time. Thirty percent more time as estimated by one mathematical biologist—who nevertheless suggests that achieving this valued (and valuable) scientific commodity is not insurmountable.

> Reproducibility is like brushing your teeth. . . . It is good for you [and] . . . once you learn it, it becomes a habit. (Irakli Loladze as quoted by Monya Baker, 2016, p. 454)

While 30% more time and effort is probably an overestimate, there will also surely be other reproducibility edicts and strategies suggested in future that will add time and effort to the research process. Some of these have already been proposed, such as "multiverse analyses," and these will most likely fall by the wayside since they require too much time and effort and actually violate some well-established practices such as employing only the most defensible and discipline-accepted procedures.

Undoubtedly, if the reproducibility initiative persists, new facilitative roles and the expansion of existing ones will also be adopted. And almost certainly, if the prevalence of QRPs and misconduct persists (or increases), some of these will gain traction such as (a) required use of institutionally centralized drives for storing all published data *and* supporting documentation (hopefully including workflows) or (b) institutions requiring the perusal of papers by an independent scientist prior to submission, with an eye toward spotting abnormalities. Existing examples of the latter include Catherine Winchester's (2018) "relatively new reproducibility role" at the Cancer Research UK Beatson Institute and some institutions' use of outside firms to conduct reproducibility screening in response to one or more egregiously fraudulent incidents (Abbott, 2019).

Perhaps even the increased awareness that a cadre of scientists is searching, finding, and promulgating published examples of irreproducibility results may encourage their colleagues to avoid the deleterious effects of QRPs. Meta-researchers, even with the considerable limitations of their approach, are already beginning to play a growing role in both increasing the awareness of substandard methodologies and tracking reproducibility progress over time.

And progress is being made. Iqbal, Wallach, Khoury, and colleagues (2016), for example, in analyzing a random sample of 441 biomedical journal articles published from 2000 to 2014 found a small but positive trend in the reporting of a number of reproducibility and transparency behaviors over this time interval. However, as the authors' noted, the continuance of such studies plays an important role in tracking the effects of the reproducibility initiative over time.

> By continuing to monitor these indicators in the future, it is possible to track any evidence of improvement in the design, conduct, analysis, funding, and independence of biomedical research over time. (p. 9)

Similarly, a more recent meta-scientific tracking study (Menke, Roelandse, Ozyurt, et al., 2020) similarly found small but positive gains from 1997 to either 2016 or 2019 involving six methodological indicators (e.g., randomization, blinding, power analysis) and six indicators related to the provision of sufficient information on the biological materials employed and that are essential for replication purposes (e.g., antibodies and cell lines). Unfortunately, the results were not especially impressive for several of these key indicators.

1. Blinding was mentioned in 2.9% of articles in 1997 and in only 8.6% of articles in 2016;
2. Any mention of a power analysis rose from a jaw-breaking 2.2% in 1997 to 9.9% in 2016;
3. Description/provision of the actual organisms employed (e.g., mice, human cancer cells) improved from 21.1% to 22.0% over the 10-year span; and
4. The identity of the cell lines witnessed an increase of less than 3% (36.8% to 39.3%).

Still, improvement is improvement, and all 12 behaviors are important from a reproducibility perspective: methodological rigor indicators because their presence dramatically impacts the prevalence of false-positive results; key biological materials because they are often a necessary condition for replication in this line of research.

This particular study is of special interest for two other reasons. First, it employs a QRP identification instrument (SciScore) that could be employed with certain modifications by other researchers in other disciplines. Second, the data generated could serve as an adjunct to education, best-practice guideline standards (e.g., CONSORT and ARRIVE), and publishing mandates as well as an evaluation tool thereof. The authors also provide an interesting case study of this genre of evaluation (as did the Kidwell et al. (2016) badges study discussed in Chapter 10).

It is worth repeating that the evidence presented so far suggests that education, professional guidelines, and unenforced mandatory best-practice publishing edicts, while better than nothing, are far from the most important drivers of reproducibility. That distinction belongs to journal editors and peer reviewers who are willing to enforce reproducibility behaviors,

preferably coupled with a growing number of online resources to facilitate compliance.

The Open Science Framework website is an impressive example of the latter; in addition, as discussed by Menke et al., one of the most useful resources for bench researchers is undoubtedly the Resource Identification Portal (RRID) which facilitates the reporting RRID identifiers in all published research employing in vivo resources. These identifiers are essential for the replication of many if not most such findings, and the RRID portal allows interested scientists to ascertain the specific commercial or other sources for said resources associated with their identifiers—and thus provides the capacity to obtain them.

Using antibodies as an example, Menke and colleagues found that, by 2019, 14 of the 15 journals with the highest identification rates participated in the RRID initiative. The average antibody identification rate for this 14-journal cohort was 91.7%, as compared to 43.3% of the 682 journals that had published at least 11 antibody-containing articles. Not proof-positive of the causal effects of the RRID initiative, but the fact is that the journal impact factor of the 682 journals was unrelated to this indicator (numerically the correlation was negative) and certainly buttressed by the 2016 decree by the editor of *Cell* (Marcus et al., 2016)—one of the most cited journals in all of science and definitely the most cited and prestigious journal in its area.

The decree itself was part of an innovation consisting of the Structured, Transparent, Accessible Reporting (STAR) system, which not only required the implementation of the RRID initiative but also required that the information be provided in a mandated structured Key Resources Table along with standardized section headings that "follow guidelines from the NIH Rigor and Reproducibility Initiative and [are] aligned with the ARRIVE guidelines on animal experimentation and the Center for Open Science's Guidelines for Transparency and Openness Promotion (https://cos.io/top/)" (Marcus et al., 1059).

Not surprisingly the compliance rate for reporting the requisite antibody information in the eight *Cell* journals was even higher than the other seven journals (93.6% vs. 91.7%) included in Menke et al.'s previously mentioned top 15 journals (see table 5 of the original article). If other journal editors in other disciplines were this conscientious, there is little question which of the two posited alternative futures would result from the reproducibility initiative as a whole.

Of Course, There Are Other Alternative Futures

Perhaps it is equally likely that neither of the preceding alternative futures will occur. Instead, it may be that something in between will be realized or some completely unforeseen paradigmatic sea change will manifest itself. But, as always, in deference to my virtual mentor, I will defer here.

In all fairness, however, I cannot place the blame for my reticence solely upon my mentor for the simple reason that the reproducibility field as a whole is in the process of changing so rapidly (and involves so many empirical disciplines with unique challenges and strategies to meet those challenges) that no single book could cover them all in any detail. So the story presented here is by necessity incomplete, and its ending cannot be told for an indeterminate period of time.

However, I have no hesitation in proclaiming that the reproducibility initiative represents an unquestionable *present-day success story* to be celebrated regardless of what the future holds. And we all currently owe a significant debt to the dedicated investigators, methodologists, and statisticians chronicled here. Their contributions to improving the quality and veracity of scientific inquiry over the past decade or so deserve a place of honor in the history of science itself.

References

Abbott, A. (2019). The integrity inspectors. *Nature, 575*, 430–433.

Asendorpf, J. B., Conner, M., De Fruyt, F., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119.

Baker, M. (2016). Is there a reproducibility crisis? *Nature, 533*, 452–454.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives in Psychological Science*, *7*, 600–604.

Gelman, A. (2018). The experiments are fascinating. But nobody can repeat them. *Sciences Times*. https://www.nytimes.com/2018/11/19/science/science-research-fraud-reproducibility.html

Grahe, J. E., Reifman, A., Hermann, A. D., et al. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives in Psychological Science*, *7*, 605–607.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., et al. (2016). Reproducible research practices and transparency across biomedical literature. *PLoS Biology*, e1002333.

Jekel, M., Fiedler, S., Torras, R. A., et al. (2019). How to teach open science principles in the undergraduate curriculum: The Hagen Cumulative Science Project. http://www.marc-jekel.de/publication/teaching_hagen/

Kidwell, M. C., Ljiljana B., Lazarević, L. B., et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PloS Biology, 14*, e1002456.

Marcus, E., for the Cell team. (2016). A STAR is born. *Cell, 166*, 1059–1060.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry, 1*, 108–141.

Menke, J., Roelandse, M., Ozyurt, B., et al. (2020). Rigor and Transparency Index, a new metric of quality for assessing biological and medical science methods. bioRxiv http://doi.org/dkg6;2020

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., et al. (2017). A manifesto for reproducible science. *Nature Human Behavior, 1*, 1–9.

Winchester, C. (2018). Give every paper a read for reproducibility. *Nature*, *557*, 281.

Index

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.

Tables and figures are indicated by *t* and *f* following the page number